UAB Universitat Autònoma de Barcelona

Student's Mental Health Clustering:

Anàlisi i clusterització de la salut mental dels estudiants mitjançant Machine Learning

Grup 05

Ilias Dahchouri - 1673164
Joan Colillas - 1670247
Bernat Vidal - 1670982

Índex

Introducció

En aquest informe es detalla el procés de planificació, desenvolupament i anàlisi dels resultats obtinguts en l'estudi que hem dut a terme.

L'estudi ha involucrat la utilització d'algorismes de machine learning, principalment algorismes de clustering, que han estat adequats per assolir l'objectiu del projecte. A més, s'han aplicat algorismes de regressió per avaluar la rellevància de les variables amb les quals treballàvem, cosa que ens permet netejar i refinar els resultats del clustering.

Metodologia

Per a fer el treball, hem seguit una metodologia basada en el plantejament setmanal de preguntes. Cada setmana, ens hem plantejat una sèrie de preguntes que ens han servit per a orientar el nostre projecte i facilitar-ne el desenvolupament. A mesura que anàvem trobant respostes a les preguntes avançàvem en l'elaboració del treball de forma organitzada.

Per a poder treballar de forma col·laboraria més fàcilment i poder avançar en el treball de forma simultània hem utilitzat GitHub, una plataforma que permet dur a terme un control de versions i gestionar els documents i codi del projecte de forma eficient. Aquesta eina ens ha permès evitar conflictes en els arxius, sincronitzar els canvis que realitza cada membre del grup i tenir un registre d'aquestes dels canvis realitzats.

Objectiu

Aquest projecte té l'objectiu d'analitzar l'estat de salut mental d'un grup d'estudiants mitjançant tècniques de Machine Learning. Utilitzarem algorismes de clusterització per agrupar els estudiants en funció de les seves característiques individuals, identificant així patrons i tendències originalment amagats que puguin ser indicatius de diferents estats mentals. Amb aquesta anàlisi, es pretén identificar quins grups d'estudiants necessiten més atenció i suport, per tal de dissenyar intervencions més eficients per millorar la seva salut mental.

Altrament, amb aquest estudi pretenem respondre aquestes dues preguntes:

- **Quins grups d'estudiants haurien de rebre major suport per a la seva salut mental?**
- **Quines característiques comparteixen els estudiants amb millor/pitjor estat de salut mental?**

Procediment per a l'estudi

La part pràctica del projecte es divideix en les següents parts:

Primerament, es neteja el dataset per assegurar-nos que les dades no contenen errors de format i/o espais en blanc.

Un cop filtrades les dades, hem de fer un estudi per saber quines variables són rellevants per al model d'aprenentatge. Per a fer-ho, utilitzarem mètodes de 'feature importance' que ens permetran conèixer quines variables podem eliminar per tal d'aconseguir millors resultats i un millor rendiment.

Seguidament, separarem les variables psicològiques objectives del dataset original, ja que aquestes ens serviran com a referència per avaluar l'estat psicològic dels usuaris. Aquesta separació és necessària, ja que per a l'estudi no té sentit fer clustering utilitzant també les variables psicològiques, perquè l'objectiu de l'estudi és aplicar el clustering en funció de característiques que no estiguin relacionades directament amb factors psicològics i comprovar després del clustering si aquestes estan relacionades amb un millor o pitjor estat de salut mental.

Un cop obtingut el dataset final, aplicarem diferents algorismes de clustering, i finalment visualitzarem i analitzarem els resultats.

Dataset

Després d'haver valorat diferents datasets trobats, vàrem optar per utilitzar el dataset que se'ns oferia en l'enunciat del projecte, el dataset "Medical Student Health Analysi" fet per Fares SAYADI.

[https://www.kaggle.com/code/faressayadi/medical-student-health-analysis-fares-sayadi/input]

Aquest recull un conjunt de dades personals dels estudiants que descriuen diversos factors relacionats amb el seu estat mental, com el seu nivell d'ansietat o de depressió, així com físics o relacionats amb la seva rutina com l'any acadèmic, si tenen o no parella, si treballen, etc.

En concret, les variables dins del dataset són les següents (classificades en variables categòriques i numèriques):

Categòriques:

- **sex:** gènere [1 -> Home | 2 -> Dona | 3 -> No binari]
- **year**: any acadèmic [1 -> Bmed1 | 2 -> Bmed2 | 3 -> Bmed3 | 4 -> Mmed1 | 5 -> Mmed2 | 6 -> Mmed3]
- **part**: parella [1 -> Si | 0 -> No]
- **job:** treball [1 -> Si | 0 -> No]
- **psyt:** s'ha consultat un psicòleg en els últims 12 mesos? [0 -> No | 1 -> Si]
- **glang:** llengua materna [1 -> Francès | 15 -> Alemany | 53 -> Català...]
- **health:** nivell de satisfacció amb la salut [1 -> Molt Baix | 5 -> Molt alt]
- **stud_h:** hores d'estudi per setmana

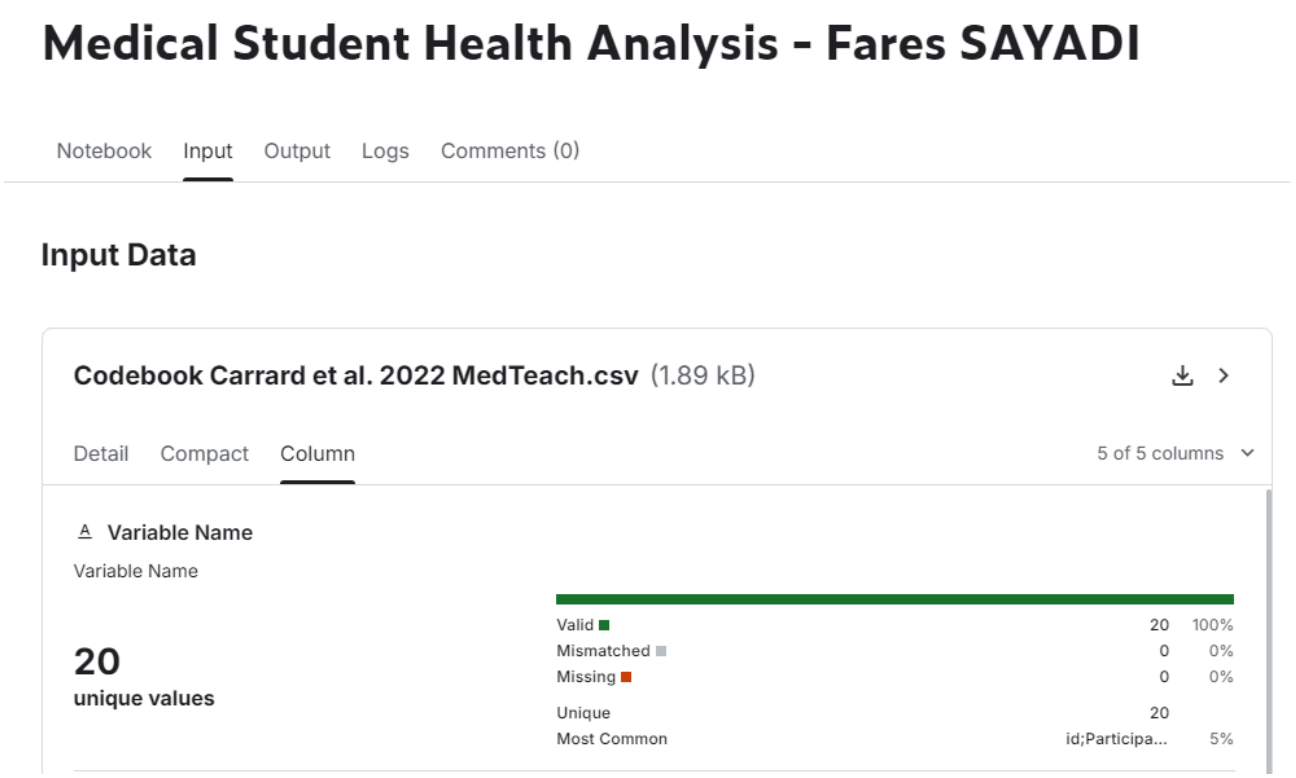
Numèriques: *[min - max]*

- **age:** edat [17 - 49]
- **jspe:** nivell d'emapatía [67 - 125]
- **qcae_cog:** nivell cognitiu [37 - 76]
- **qcae_aff:** nivell d'afecció [18 - 48]
- **erec_mean:** percentatge total de respostes correctes en el GERT, un test on s'avalua si les persones poden reconèixer les emocions basant-se en llenguatge no verbal [0.35 - 0.95]
- **cesd:** escala de depressió [0 - 56]
- **stai_t:** escala d'ansietat [20 - 77]
- **mbi_ex:** cansament emocional [5- 30]
- **mbi_cy:** cinisme -> Mesura que tan distant una persona se sent respecte al seu voltant [4 - 24]
- **mbi_ea:** eficàcia acadèmica [10 - 36]

Addicionalment, el dataset també inclou dues altres variables: id i amsp, les quals s'han descartat inicialment.

Pel que fa a les dimensions del dataset, originalment compta amb 886 registres (files) i 20 variables (columnes).

Capçalera del dataset de kaggle:



Cleaning/Preprocessament

Com ja hem dit, el dataset conté 886 registres i 20 columnes. El preprocessament va ser fet en part utilitzant l'script del únic starting point que hem fet servir. En ell s'eliminaven els registres que no tinguessin valors en alguna de les seves columnes, les dades repetides etc. No obstant, després del procés el número de registres la mida del dataset no va variar ja que les dades ja estaven processades i filtrades des d'un principi (en el kaggle ja estaven processades). En quant a les columnes, varem eliminar 2 columnes del dataset en aquesta fase:

- **id:** no te rellevància en el clustering
- **amsp:** no està document i no hi ha informació de què representa la variable, impossibilitant l'anàlisi de cap resultat referent a aquest.

El dataset conté variables de diferents tipus, les quals han estat dividides en grups segons la seva categoria i tractades segons les seves característiques.

En primer lloc, hem eliminat les variables psicològiques objectiu del dataframe original per evitar que influïssin en el procés de clústering, ja que el nostre objectiu és agrupar els estudiants en funció de factors no directament relacionats amb l'estat psicològic.

Les variables binàries, com part (parella) o psyb (si s'ha visitat un psicòleg en l'últim any), no han estat modificades, ja que ja són adequades per a la seva utilització en el model de clústering.

Les variables categòriques amb ordre (ordinals) i les variables numèriques s'han normalitzat i estandarditzat utilitzant el StandardScaler per assegurar-nos que totes tenen la mateixa escala, evitant que algunes variables tinguin més rellevància en el procés de clústering només per tenir magnituds més grans. Aquest pas era essencial ja que el clústering es basa en distàncies entre punts, i la presència de valors en escales diferents podria distorsionar els resultats.

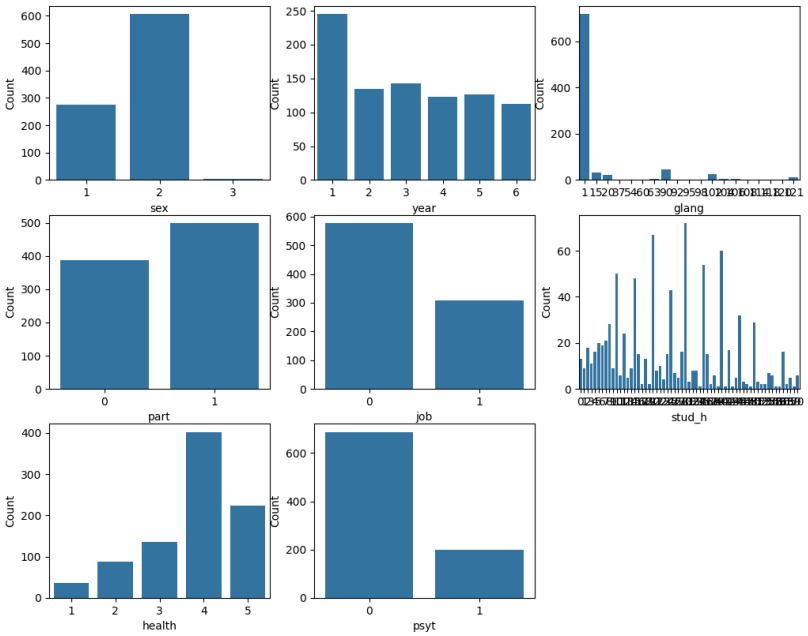
Les variables categòriques sense ordre (nominals) s'han dividit en múltiples columnes mitjançant el one-hot encoding, creant tantes variables com categories existissin en la variable nominal. Aquest pas també era essencial ja que els algorismes de clústering no poden gestionar variables categòriques de manera directa. Aquesta transformació permet representar les categories en un format numèric adequat per al model, assegurant que cada valor es tracti de manera independent en el procés d'agrupament.

Starting Point

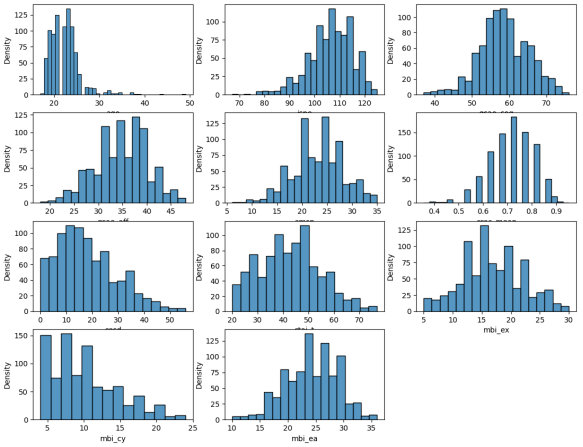
Com a starting point hem utilitzat el codi que venia inclós en la pàgina del dataset de kaggle. Aquest consta d'un notebook (python .ipynb) en el qual es tracten diferents punts.

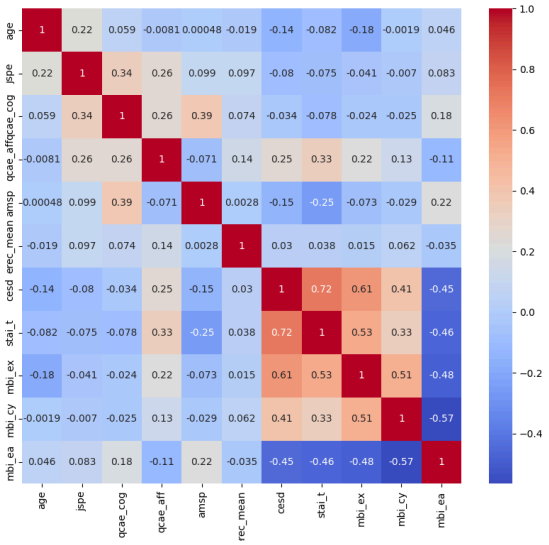
En primer lloc, inclou un model de preparació i filtratge inicial de les dades. En concret, és comprova que no hi ha registres duplicats ni registres amb cap dels seus valors en blanc o amb valors nuls. Es divideix el dataset en dos grups: dataset amb variables categòriques i un d'altre amb les variables numèriques. Finalment, s'elimina la columna 'id' ja que no té rellevància.

En segon lloc, es fa un anàlisi de les variables categòriques del dataset, en el qual s'utilitzen gràfics de barres per mostrar les distribucions de les categories i es fa una prova de chi-quadrat per avaluar la independència entre parelles de variables categòriques.

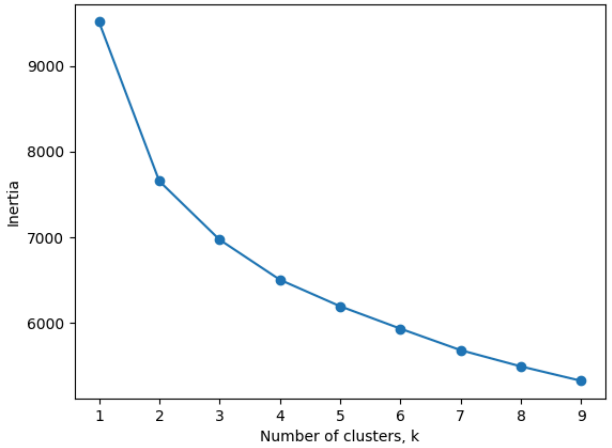


Seguidament, es fa un anàlisi de les dades numèriques, en el qual es mostren les seves distribucions, es calcula la matriu de correlació entre les variables i es representa amb un heatmap. Finalment, s'aplica pca per reduir la dimensionalitat de les variables per utilitzar-se posteriorment en els clusterings.





Després d'analitzar i transformar les dades (amb pca) es procedeix a fer el clústering. En aquest cas només s'aplica l'algorisme de K-means per a fer els clusterings. Alhora, es busca el nombre òptim de clusters utilitzant la inèrcia i creant gràfics de dispersió per representar-lo. Es guarden els resultats dels clusters al dataframe original creant una nova columna per guardar els labels i finalment es mostren les propietats dels clusters mitjançant valors mitjans.

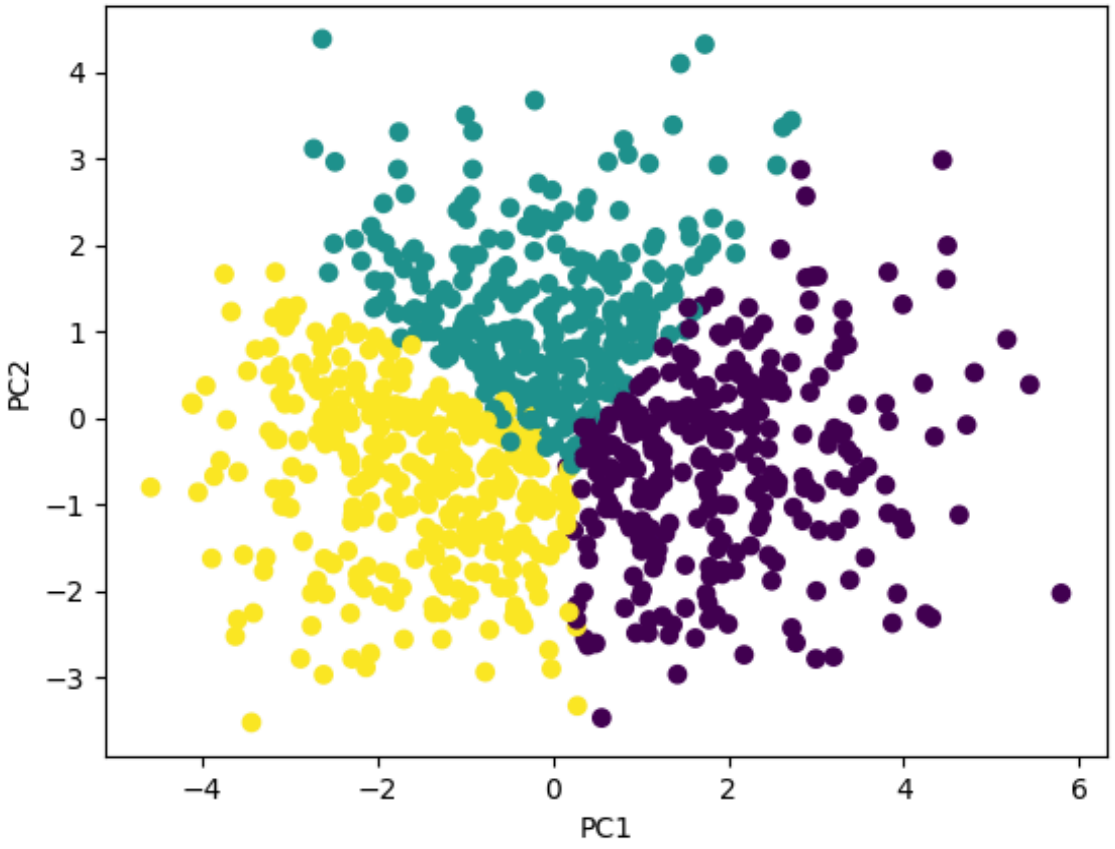


En l'últim pas s'entrena un model de regressió logística per predir quins serien els labels dels clusters. Després, es mostra la matriu de confusió i l'informe de classificació per avaluar l rendiment del model.

El 'starting point ha sigut útil ja que ens ha donat una idea de com podria ser el nostre codi final. No obstant, vàrem detectar que hi havia parts que contenien errors o que no s'adaptaven al nostre objectiu.

En quant al filtratge de dades, no hi vàrem detectar problemes, i per tant, es va fer servir per netejar el dataset, també vàrem utilitzar les funcions de representació de distribucions per analitzar el dataset.

No obstant, vàrem trobar que no tenia sentit utilitzar el dataset amb reducció de dimensionalitat (pca) per a fer el clustering, ja que d'aquesta manera estàvem perdent precisió i detall a la hora de fer el clústering, ja que d'aquesta manera estàvem perdent precisió degut a que no s'estava fent el clústering amb totes les variables. No obstant això, la representació visual sí seria millor i es veurien els clústers més diferenciats.



Feature importance i models de classificació binària

Des d'un inici, podem explicar el que vam veure en el starting point extraient observacions rellevants per a entendre entre altres qüestions perquè quan fèiem clusterings patíem molts solapaments i no eren entendibles:

Distribució desequilibrada:

- Per a la variable *sex*, la gran majoria dels participants són del gènere femení (68,39%), seguit dels masculins (31,04%), amb només un 0,56% d'individus no binaris. Aquest desequilibri pot generar un biaix en la rellevància percebuda de *sex*, quan en realitat ens interessen característiques de l'estudiant com a tal i no tant el seu gènere a l'hora d'ajudar-lo.
- Per a la variable *glang*, el 80,93% dels participants comparteixen el mateix idioma nadiu. La resta d'idiomes estan distribuïts de manera molt desigual, cosa que dificulta detectar patrons significatius que no siguin simplement un reflex d'aquesta predominança.

Ens van sorgir les següents preguntes:

- El gènere i la llengua materna (glang) són realment variables predictives de l'estat mental o només són un reflex de biaixos inherents al conjunt de dades?
- Si conservem el gènere i glang com a variables clau, correm el risc de construir models que reflecteixin més estereotips que relacions reals?
- Per ajudar als estudiants segons l'estat mental, els nostres models utilitzaran variables interessants per predir o es veuran esbiaixades per variables com sex i glang que conceptualment no ens ajuden a predir l'estat de salut mental de l'individu més enllà d'estereotips.

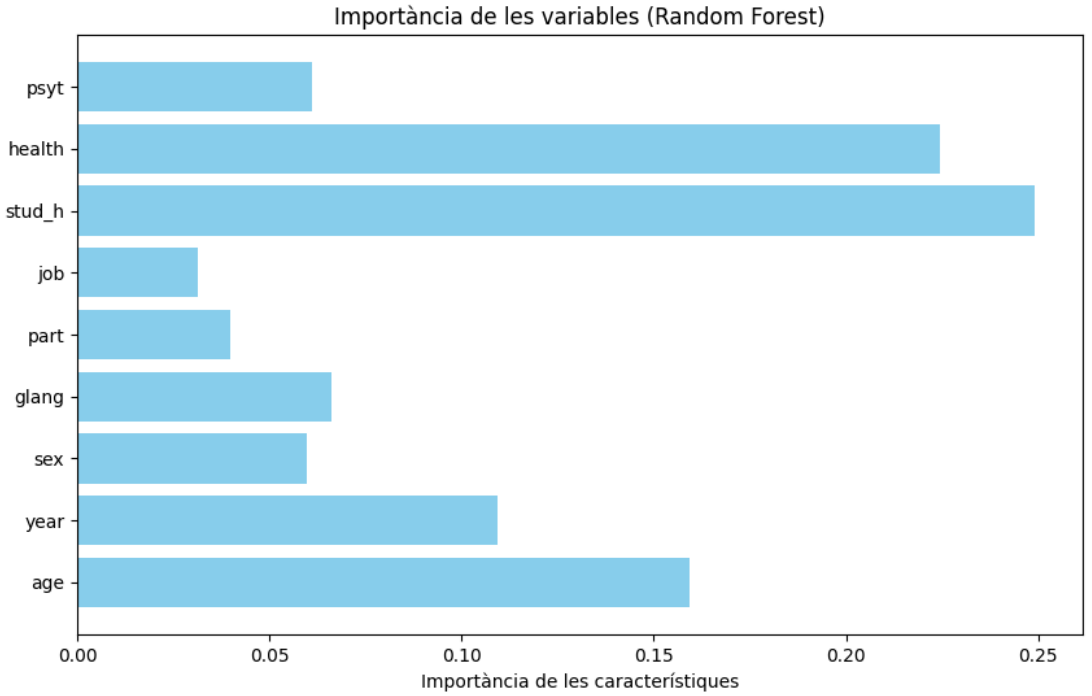
Chi-square test:

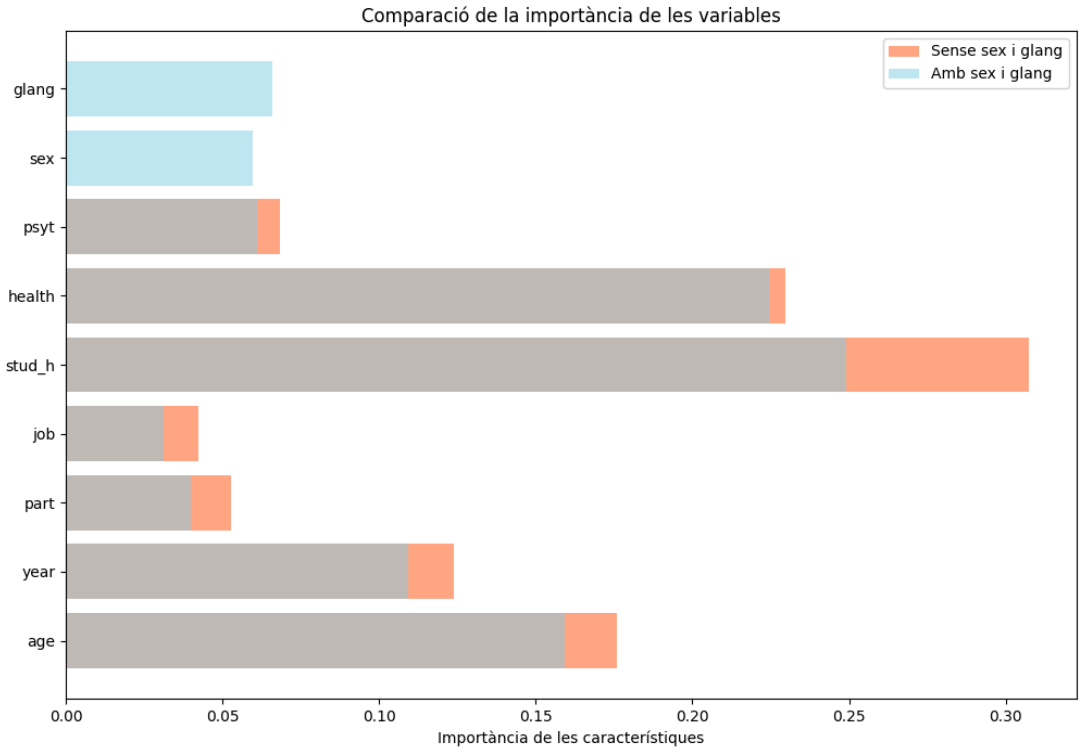
- El *chi-square test* mostra una significació estadística entre algunes variables, com *sex* i *psyt*, o *glang* i *health*. Tot i això, aquestes relacions poden no tenir impacte causal en les variables psicològiques (com *cesd*, *stai_t*, o *mbi_ex*) si les seves contribucions al model predictiu són baixes.

Distribució visual i solapaments:

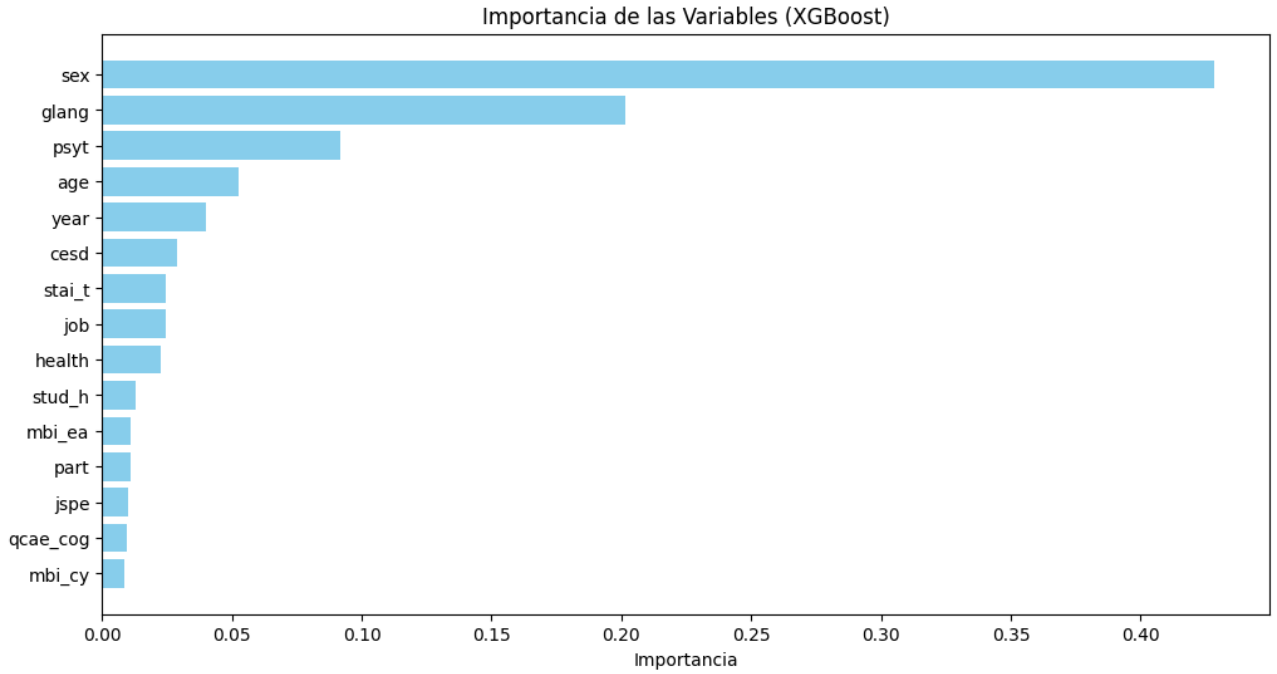
- Les distribucions dels valors de *sex* i *glang* en relació amb altres variables presenten solapaments significatius i no mostren diferències clares entre categories per predir l'estat de salut mental o classificar individus en clústers de salut mental que vam fer inicialment.

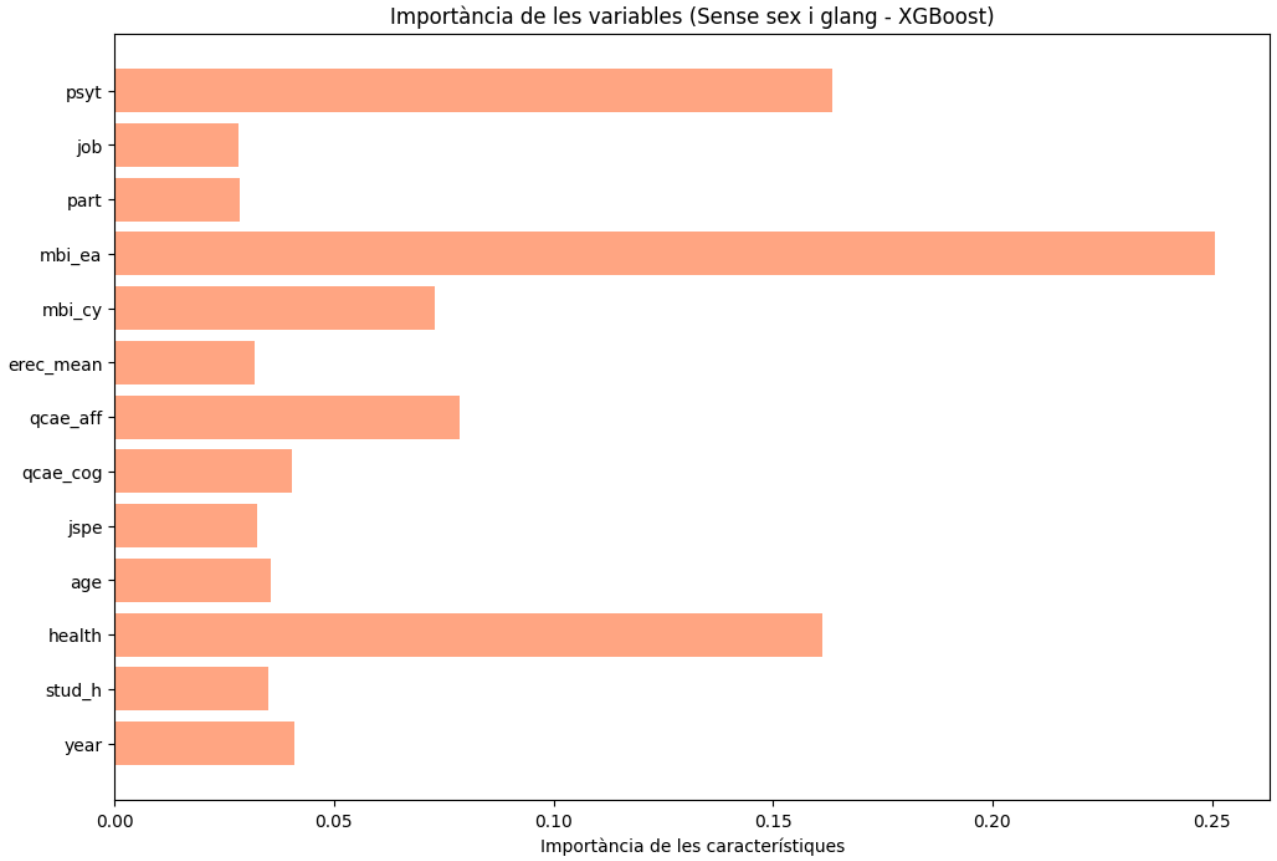
RandomForest Regressor





XGBoost Regressor





Com i que hem fet per arribar a visualitzar aquestes gràfiques?

RandomForest Regressor:

1. **Eix Y (importància de les variables)**
    - L'eix Y mostra les variables que formen part de l'anàlisi (per exemple, stud_h, health, psyt, etc.).
    - Aquestes variables són les característiques (features) que l'algorisme analitza per determinar la seva rellevància en relació amb la predicció de les variables psicològiques (cesd, stai_t, mbi_ex).
2. **Eix X (valors d'importància)**
    - Els valors numèrics a l'eix X representen la importància relativa de cada variable.
    - En el cas del Random Forest aquesta importància s'obté calculant quantes vegades i amb quina rellevància cada variable contribueix a dividir correctament les dades en els arbres de decisió.
    - Els valors van de 0.0 a 0.30 o 0.25 segons el gràfic, i reflecteixen el pes percentual de cada variable respecte al total.
        - Exemple: Si una variable té un valor de 0.25, significa que té una contribució relativa del 25% en el procés de predicció.
3. **Diferència entre les dues gràfiques**
    - **Primera gràfica:** Compara les importàncies amb sex i glang.
    - **Segona gràfica:** Mostra la importància de totes les variables amb sex i glang i sense aquestes dues variables.
    - Es pot observar que sex i glang tenen valors molt baixos en ambdós casos, cosa que indica que no són determinants per predir les variables psicològiques més significatives, a més que són les que provoquen més soroll en les nostres anàlisis.

XGBoost:

En el cas de XGBoost, els gràfics mostren la importància de les variables utilitzades en l'anàlisi.

1. **Eix Y (importància de les variables)**
   Aquest eix inclou totes les característiques analitzades, com ara stud_h, health, psyt, etc, juntament amb les variables psicològiques més rellevants com cesd, stai_t i mbi_ex. Aquestes variables són avaluades per determinar la seva contribució relativa a la predicció del nostre target (estat de salut mental).
2. **Eix X (valors d'importància)**
   L'eix X indica la rellevància de cada variable en el model XGBoost, que es calcula basant-se en la freqüència i la influència de cada característica en la millora del resultat en el model de boosting. Els valors d'aquest eix van de 0.0 a un màxim (en el cas de les nostres gràfiques, aproximadament 0.25-0.30), i mostren el pes relatiu de cada variable en el conjunt total d'importància.
3. **Observacions dels gràfics**

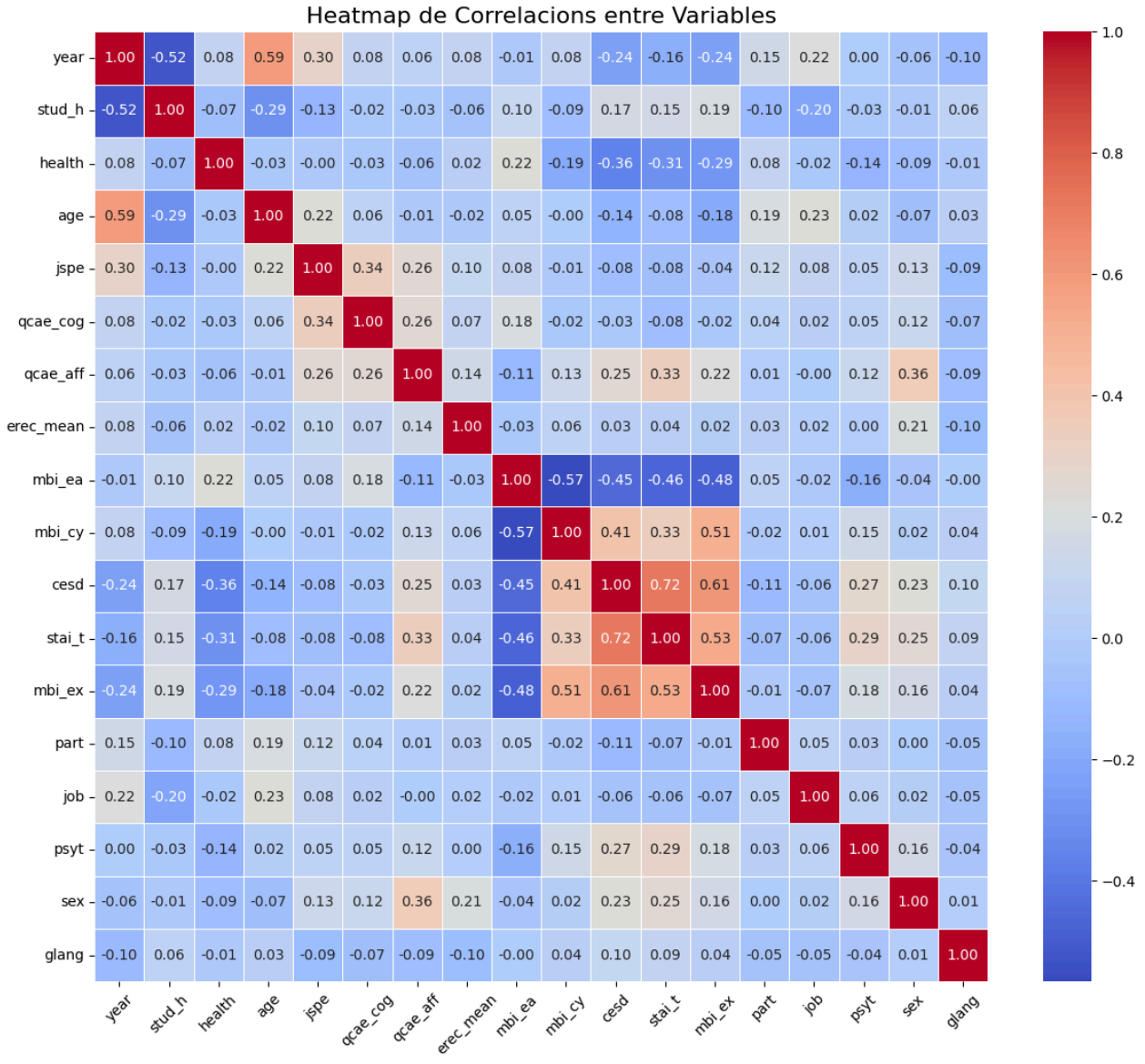
- En la primera gràfica, observem com sex i glang tenen una importància aparentment molt elevada. Això s'explica perquè XGBoost tendeix a assignar importància a variables amb distribucions desequilibrades (com la majoria d'estudiants assignats a sex=2 i glang=1), tot i que aquestes no tinguin una correlació significativa amb les variables psicològiques, tal com veurem en els heatmaps de correlacions.
- En canvi, en la segona gràfica, que exclou sex i glang, les variables com mbi_ea, health i psyt emergeixen com les més significatives, reflectint millor la influència real de les

variables en la predicció de l'estat psicològic, coneixen que cesd, stai_t i mbi_ex són les variables psicològiques més importants. Aquesta eliminació permet reduir el soroll introduït per variables menys rellevants.

En resum, els resultats obtinguts amb XGBoost ens permeten confirmar que les variables cesd, stai_t i mbi_ex, juntament amb factors com health i mbi_ea, són les més determinants per explicar l'estat de salut mental. L'aparent importància de sex i glang en la primera gràfica demostra com certes variables poden crear biaixos i soroll, destacant la necessitat de considerar tant l'anàlisi de correlacions com el coneixement del domini per interpretar correctament els resultats.

Justificació de les variables significatives en clustering o classificació

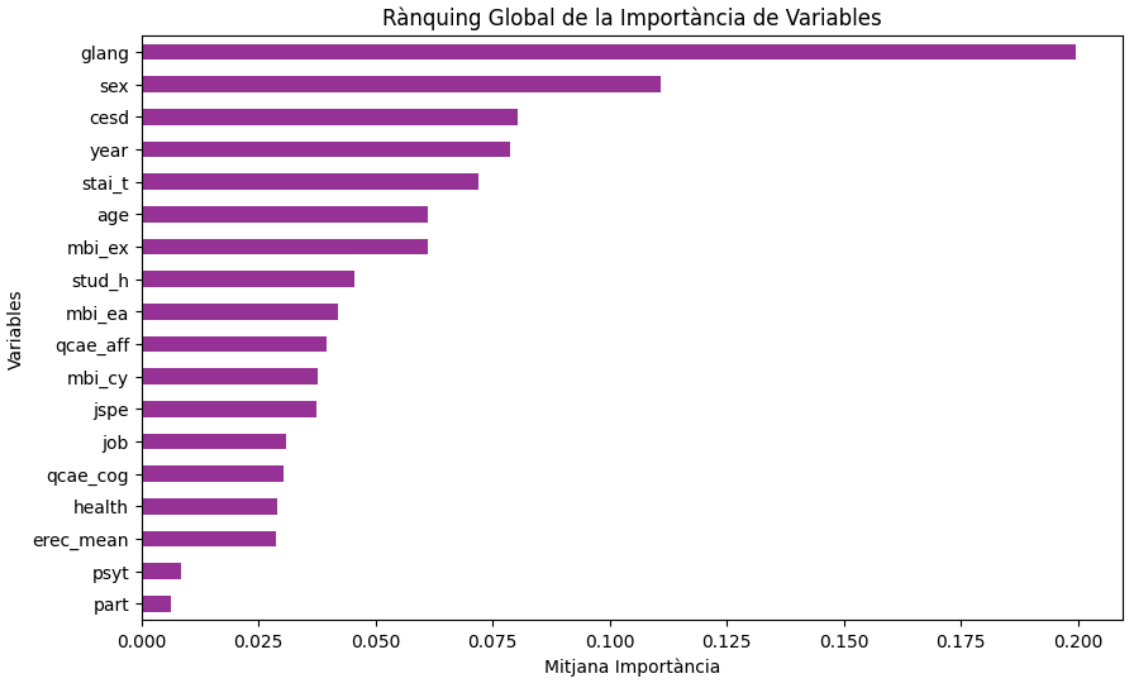
Heatmap amb correlacions entre totes les variables correctament preprocessades:

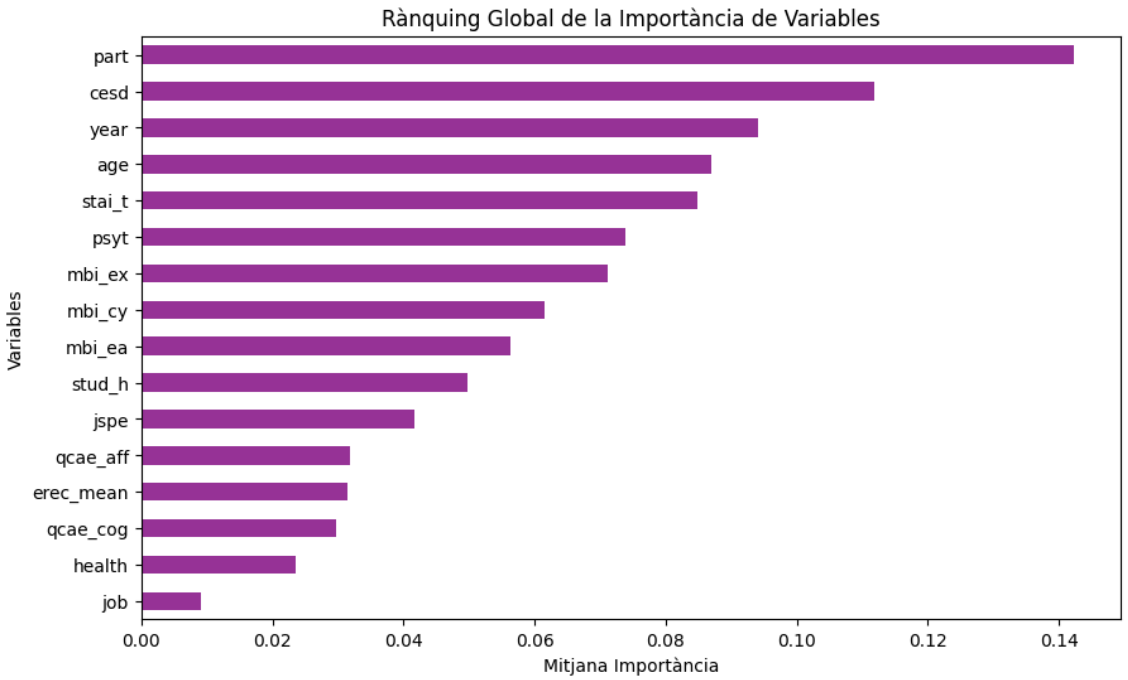


El heatmap de correlacions proporciona una visió clara de la relació entre les variables mitjançant mesures estadístiques, destacant que CESD, STAI_T i MBI_EX tenen una forta correlació positiva entre elles, amb valors que superen el 0.7. Aquesta forta associació indica que comparteixen una gran part de la informació i expliquen una proporció significativa de la variabilitat psicològica en el

conjunt de dades. Aquestes correlacions suggereixen que els patrons observats en aquestes tres variables són indicadors clau de l'estat mental dels estudiants i, per tant, les converteixen en els targets principals del projecte per categoritzar la salut mental en nivells bons o dolents. D'altra banda, variables com sex i glang mostren correlacions baixes amb CESD, STAI_T i MBI_EX, evidenciant que no contribueixen de manera significativa a explicar la variabilitat psicològica ni tenen relació directa amb l'estat mental. Tot i que el heatmap és una eina estadística, reforça els resultats observats en tècniques de machine learning, on CESD, STAI_T i MBI_EX també han demostrat ser dominants en els models predictius i de clustering. Això valida que la nostra selecció de variables i l'enfocament del projecte són adequats, tant des d'una perspectiva estadística com des de l'anàlisi de dades amb machine learning.

Importància de variables a l'hora de fer els clusterings:





L'objectiu principal del codi que ens permet visualitzar les gràfiques anteriors és avaluar la importància de les variables en el procés de clustering per determinar si algunes variables expliquen de manera significativa l'agrupament dels individus. Per aconseguir-ho, les dades han estat preprocessades per garantir que totes les variables, incloent-hi les categòriques com sex i glang, tinguin el mateix pes inicial, evitant així biaixos a causa d'escales diferents. Posteriorment, hem aplicat algoritmes de clustering com KMeans, GMM i Agglomerative Clustering, seleccionant el nombre òptim de clusters amb criteris específics: Elbow Method per a KMeans, BIC per a GMM i Silhouette Score per a Agglomerative. Amb un Random Forest Classifier, hem analitzat la importància relativa de les variables en la formació dels clusters, presentant els resultats mitjançant gràfics de barres. Finalment, hem calculat un rànquing global de variables basat en la mitjana de les importàncies en els tres algoritmes, identificant així les variables més rellevants per a tot el projecte.

No obstant això, hem observat que en el cas del clustering amb GMM, les variables sex i glang han estat assignades una importància molt elevada, la qual cosa ha influït en el rànquing global, situant-les com les més destacades. Aquest resultat contrasta amb les anàlisis prèvies realitzades mitjançant heatmaps, correlacions i estudis d'importància de variables amb Random Forest Regressor, on aquestes dues variables no han aparegut com a predominants. La raó d'aquest comportament és que GMM i, en alguns casos, XGBoost atribueixen més rellevància a variables amb valors majoritaris en el dataset, com és el cas de sex i glang, on un valor concret (per exemple, sex=2 o glang=1) domina àmpliament. Aquest fet subratlla la necessitat d'eliminar aquestes variables que no aporten informació rellevant per a respondre les preguntes inicials, a més amb l'eliminació podem entendre millor les relacions entre les variables psicològiques i no psicològiques.

A més, les variables psicològiques més significatives, **cesd, stai_t i mbi_ex**, són essencials en aquest projecte per diverses raons. En primer lloc, estan altament correlacionades entre elles, com s'observa en heatmaps i matrius de correlació, reflectint una forta relació lineal i no lineal que explica l'estat mental dels individus. A més, aquestes variables són dominants en el clustering, ja que tendeixen a formar agrupacions més diferenciades que reflecteixen nivells de salut mental. Finalment, són els targets principals, perquè la seva predicció ens permet categoritzar els

estudiants en grups de salut mental bo o dolent, alineant-se amb l'objectiu inicial. Alhora, variables no psicològiques com year (curs acadèmic), psyt(s'ha consultat un psicòleg en els últims 12 mesos?), age i stud_h (hores d'estudi) també aporten informació útil, influint indirectament en l'estat psicològic. Finalment, els models van guanyar precisió i van ser capaços de formar clústers més coherents i explicatius. Això va permetre centrar-nos en patrons que realment diferencien l'estat de salut mental dels estudiants, basant-nos exclusivament en variables rellevants, reafirmant que els nostres models es basen en identificar patrons subjacents que siguin representatius de la realitat, no en perpetuar biaixos estructurals presents al conjunt de dades.

Així doncs, el dataframe final sense les variables psicològiques objectiu i amb les corresponents transformacions tindria unes dimensions de 886 x 13, essent les columnes:
'year', 'stud_h', 'health', 'age', 'jspe', 'qcae_cog', 'qcae_aff', 'erec_mean', 'mbi_ea', 'part', 'job', 'psyt', 'mbi_cy'

Podem assumir que sí que tenim prous dades per a fer el clustering. Tenim 886 registres i 13 variables en el dataframe final, aconseguint una relació de 886 / 13 = 68 registres/variable el qual és un valor acceptable que ens permetrà assegurar-nos que no hi haurà problemes d'overfitting i que els patrons trobats pel model de clustering seran significatius i representatius de les dades.

Classificadors binaris

L'avaluació de la importància de les variables, el que denominem com a 'feature importance' va jugar un paper fonamental en la nostra metodologia, guiant-nos des dels intents inicials de clustering cap a un enfocament més directe amb models de classificació binària. Aquesta decisió es va basar en diversos factors que vàrem explicar prèviament al feature importance.

Davant la complexitat i solapament dels clusters en les dades inicials, vam decidir simplificar l'anàlisi per a identificar patrons més clars amb un enfocament classificador binari, el qual ens va permetre:

- **Definir estats clars de salut mental:** Mitjançant un **llindar basat en el "mental health index"** (mitjana de CES-D, STAI-T, i MBI-EX), vam dividir els estudiants en dos grups: aquells amb malestar psicològic significatiu (necessiten més ajuda) i aquells amb millor salut mental (necessiten menys ajuda).
- **Evitar soroll en els resultats:** Eliminant la dependència de clusters difícils d'interpretar, vam aconseguir resultats més directes i pràctics, també en part gràcies a l'estudi previ del feature importance.
- **Facilitar conclusions accionables:** Amb la classificació binària, vam poder determinar amb més claredat quins estudiants necessiten suport.

Regressió logística

En el context del nostre projecte, hem emprat la regressió logística per abordar la tasca de predir l'estat de salut mental dels estudiants basant-nos en variables tant psicològiques com no psicològiques. Aquesta elecció no només ens ha permès analitzar com les variables psicològiques més significatives contribueixen a determinar l'estat de salut mental, sinó que també ha obert la possibilitat d'explorar si les variables no psicològiques podrien actuar com a predictors indirectes d'aquestes variables clau.

Si les variables no psicològiques demostren tenir una capacitat predictiva significativa sobre les variables psicològiques més rellevants (CES-D, STAI-T i MBI-EX), estaríem establint una relació que simplificaria el nostre model, reduint el nombre de variables a considerar i alhora mantenint l'objectiu final: diferenciar entre estudiants amb un estat de salut mental bo o dolent. Aquest enfocament redueix la complexitat del model, mantenint la capacitat explicativa i predictiva, fet que és essencial en projectes com el nostre on la interpretabilitat és prioritària.

Les variables psicològiques seleccionades (CES-D, STAI-T i MBI-EX) no només són altament correlacionades entre elles, sinó que també s'han demostrat com les més determinants en la diferenciació d'estats de salut mental. Això s'ha constatat a través d'anàlisis preliminars com el clustering, on aquestes variables han tingut un pes dominant en la formació dels grups, i mitjançant tècniques com XGBoost, que han destacat aquestes variables com les de major importància.

D'altra banda, si aconseguim demostrar que les variables no psicològiques poden predir amb precisió aquestes variables psicològiques clau, això tindria implicacions significatives. En primer lloc, indicaria que és possible fer una avaluació inicial de l'estat de salut mental d'un estudiant sense necessitat d'obtenir dades directament psicològiques, la qual cosa seria útil en contextos en què aquestes dades poden ser difícils d'obtenir o poden estar subjectes a biaixos personals. En segon lloc, aquesta simplificació obriria la porta a models més senzills i eficients que, malgrat això, serien capaços de captar els patrons crítics necessaris per predir l'estat de salut mental.

En definitiva, l'anàlisi i la selecció de variables no són només una qüestió tècnica, sinó que estan profundament alineades amb l'objectiu central del projecte: utilitzar la informació disponible per proporcionar eines interpretables i fiables per entendre i predir l'estat de salut mental dels estudiants. Aquest procés ens permet diferenciar de manera efectiva aquells estudiants que podrien necessitar un suport psicològic més intens d'aquells que es troben en una millor situació, establint així una base robusta per a futures intervencions i anàlisis.

Com ho hem implementat en el nostre projecte?

Per a entrenar el model, hem considerat dos conjunts de variables:

- **Variables psicològiques més importants (target):** CES-D (depressió), STAI-T (ansietat), MBI-EX (burnout emocional). Aquestes variables estan fortament correlacionades i són indicadors directes de la salut mental.
- **Variables no psicològiques i psicològiques menys importants:** Variables acadèmiques (any d'estudi, salut general) i demogràfiques (edat, gènere). Aquestes són menys correlacionades amb la salut mental, però poden oferir informació complementària.

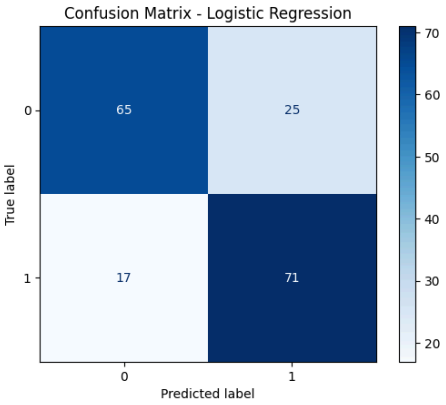
Abans d'entrenar el model, hem realitzat els passos següents, realment a tots els models hem seguit els mateixos passos:

1. **Normalització i estandardització:**
   Les variables d'entrada han estat escalades per garantir que tinguin una distribució comparable, evitant que una variable amb una escala més gran domini el model.

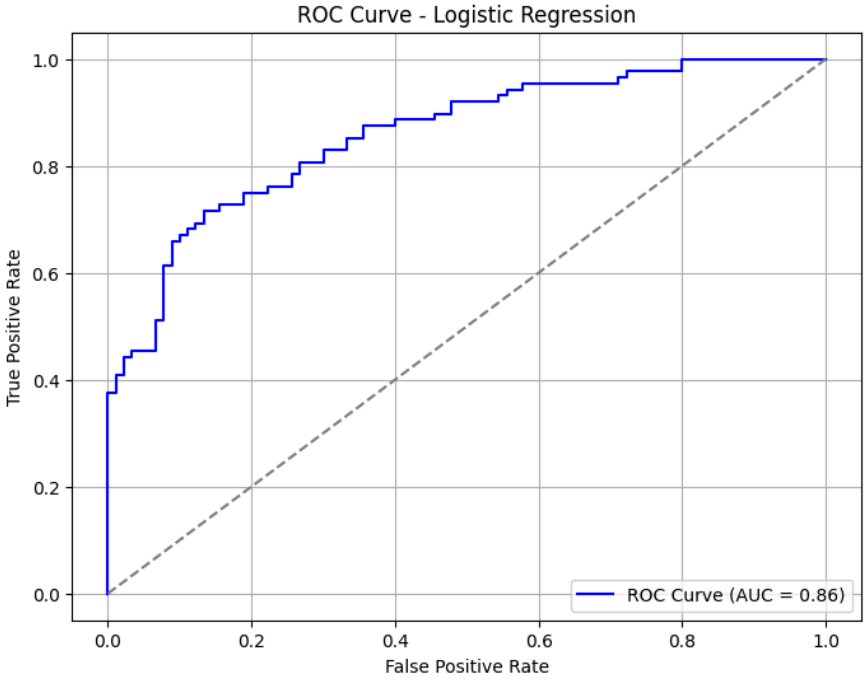
2. **Split entre entrenament i test:**
   Hem dividit el dataset en un conjunt d'entrenament (80%) i un conjunt de test (20%) per avaluar la generalització del model.

3. **Transformació de les etiquetes:**
   La variable objectiva (estat de salut mental) s'ha binaritzat en 1 (mal estat) i 0 (bon estat) basant-nos en un llindar establert amb l'índex de salut mental.

Les mètriques que hem analitzat inclouen:



- **Accuracy:** Proporció de prediccions correctes sobre el total. Ens dona una visió general del rendiment del model.
- **ROC Curve i AUC:** La corba ROC mostra la capacitat del model de distingir entre les dues classes, i l'AUC ens dona una mètrica quantitativa del rendiment (valors propers a 1 indiquen un model excel·lent).
- **Classification Report:** Metriques com precision, recall i F1-score ens ajuden a entendre el balanç entre falsos positius i falsos negatius.



Random Forest

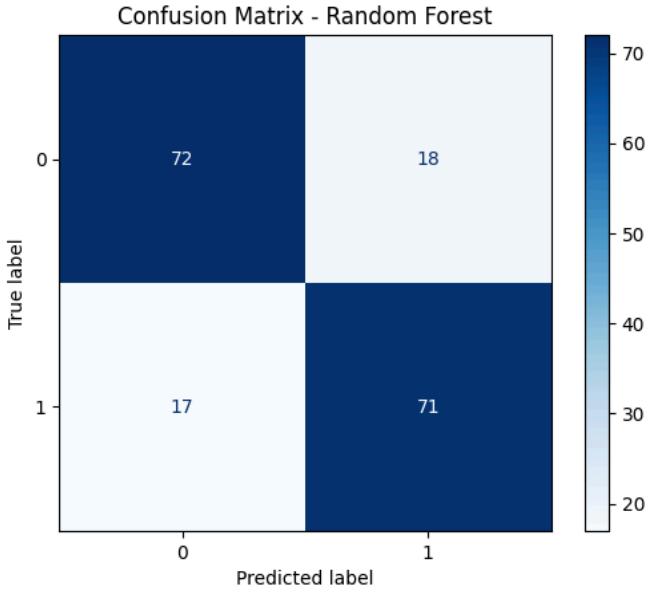
En el context del nostre projecte, hem utilitzat el *Random Forest* per abordar tant la classificació binària de l'estat de salut mental com l'anàlisi d'importància de les variables, per entendre millor també les relacions complexes entre les variables que hem comentat prèviament. Aquí desglossem tots els aspectes rellevants d'aquest model, la seva aplicació en el nostre treball i com complementa altres models com la regressió logística.

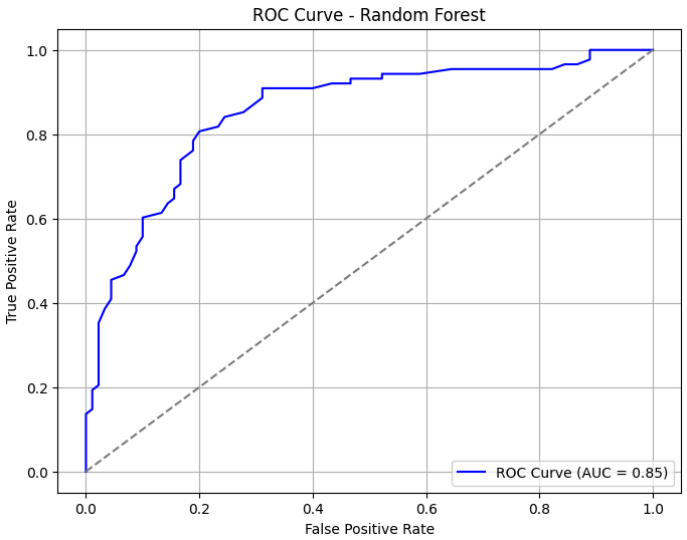
Com ho hem aplicat en el nostre model?

Hem utilitzat el *Random Forest* per predir si un estudiant es troba en un estat de salut mental bo o dolent, basant-nos en variables tant psicològiques com no psicològiques. Les seves capacitats d'identificar relacions no lineals i interaccions complexes han estat claus per millorar la precisió de les nostres prediccions respecte a models més simples com la regressió logística.

El *Random Forest* ens permet extreure la importància relativa de cada variable, mesurant com contribueixen a reduir l'error de predicció al llarg dels arbres. Aquesta funcionalitat ha estat crucial per:

- Determinar que les variables psicològiques CESD (depressió), STAI-T (ansietat) i MBI-EX (esgotament emocional) són les més rellevants per predir l'estat de salut mental.
- Justificar perquè les variables no psicològiques poden ajudar a predir, indirectament, les condicions psicològiques més importants.
- Identificar el biaix introduït per certes variables, com el gènere, que en alguns models apareix com a predominant, però que no aporta informació real significativa.





Per què Random Forest?

El Random Forest és especialment adequat per al nostre projecte perquè pot capturar relacions no lineals i interaccions complexes entre variables psicològiques i no psicològiques, cosa que la

regressió logística no pot fer per la seva limitació a relacions lineals. A més, és menys sensible a la multicolinearitat i proporciona informació crucial sobre la importància de les variables, validant així les nostres decisions de preprocessament. Amb mètriques sòlides com l'accuracy, el F1-score i la ROC Curve, el Random Forest ha demostrat ser robust i fiable. Els seus avantatges inclouen la resistència al sobreajustament gràcies al bagging, la capacitat de manejar moltes característiques sense una selecció estricta prèvia i la seva interpretabilitat, que és superior a altres models més complexos com el XGBoost. Tot i que Random Forest ha ofert resultats excel·lents, l'ús complementari d'altres models com XGBoost ens ha permès abordar millor els biaixos.

XGBoost

És un dels algorismes més potents en l'aprenentatge computacional i ha estat una peça clau en el nostre projecte per abordar la problemàtica de l'importance feature. En aquest apartat, aprofundirem en les característiques, l'aplicació, els beneficis, i les limitacions d'aquest model en el nostre context, així com en les raons que justifiquen el seu ús i els resultats obtinguts.

Per què XGBoost?

El XGBoost es construeix sobre la base del *gradient boosting*, un enfocament que crea seqüències d'arbres de decisió on cada arbre posterior intenta corregir els errors del model anterior. Això li permet capturar relacions complexes, no lineals, i interaccions entre variables. Aquestes capacitats són especialment útils en el nostre projecte per les següents raons:

1. **Relacions no lineals**: Les variables psicològiques (**CESD, STAI-T, MBI-EX**) i no psicològiques (**age, health, stud_h, part, etc**) mostren patrons i correlacions difícils d'identificar amb models lineals com la regressió logística. El XGBoost, gràcies al seu enfocament iteratiu, pot modelar aquestes relacions amb gran precisió.

2. **Gestió de dades desordenades i soroll**: En l'anàlisi del feature importance, vam observar que variables com **glang** o **sex** introduïen biaixos o soroll, dificultant el clustering i la predicció. El XGBoost incorpora regularització (*L1* i *L2*), que redueix l'impacte de variables menys rellevants, prioritzant aquelles que aporten informació valuosa, tot i que com hem vist abans al feature importance predominaven aquestes dos variables sense tenir informació valuosa pel nostre objectiu.

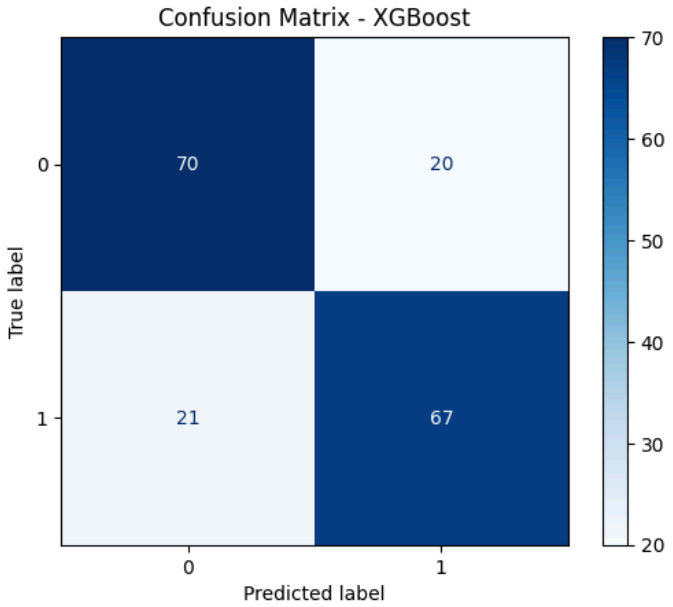
Aplicació en el Nostre Projecte

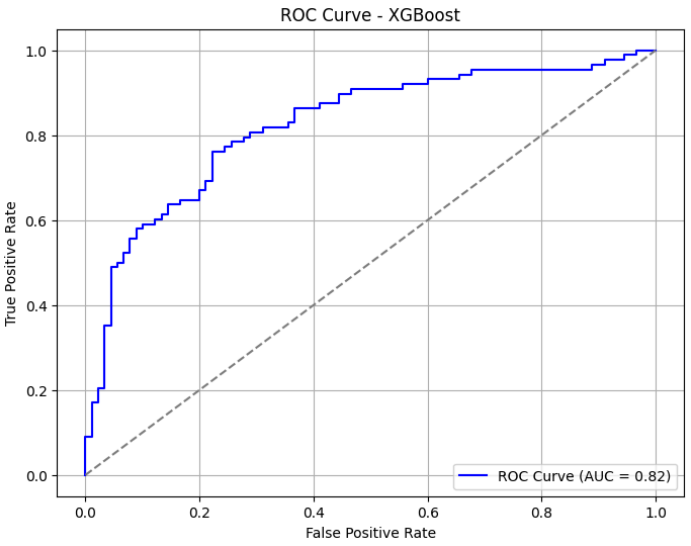
Després d'identificar variables clau i eliminar soroll, vam utilitzar XGBoost per abordar dues tasques principals:

1. **Feature Importance**: El XGBoost ens ha permès quantificar la importància relativa de cada variable en predir l'estat de salut mental. Les variables psicològiques (**CESD, STAI-T, MBI-EX**) van destacar com les més influents, cosa que valida la nostra elecció d'usar-les com a indicadors principals. També vam observar que certes variables no psicològiques, com **psyt** o **health**, tenen una contribució significativa, justificant el seu ús com a complements en la predicció.

2. **Predicció de l'estat de salut mental**: A través de la classificació binària (bé/malament psicològicament), el XGBoost ha aconseguit resultats consistents amb uns valors alts de *ROC AUC* i *accuracy*, indicant que el model capta patrons complexos en les dades.

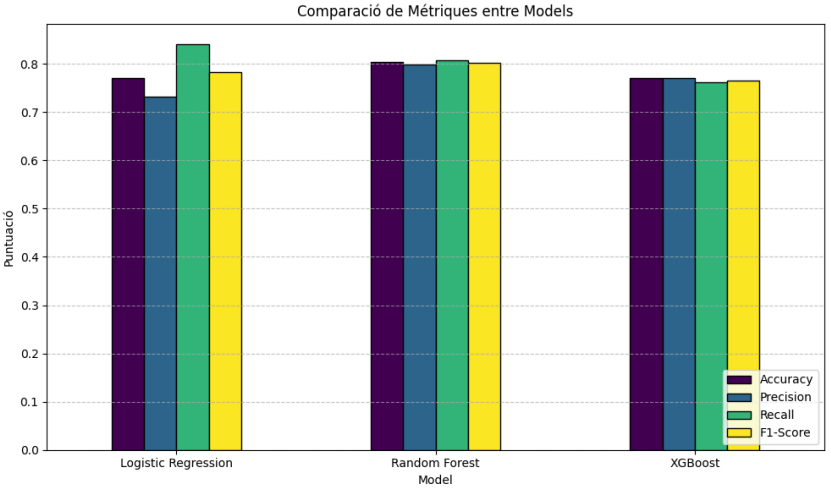
Com hem pogut visualitzar aquests resultats?





Comparació i conclusions:

	Acurracy	Precision	Recall	F1-Score
Logistic Regression	0.769663	0.732673	0.840909	0.783069
Random Forest	0.803371	0.797753	0.806818	0.802260
XGBoost	0.769663	0.770115	0.761364	0.765714



La conclusió dels resultats obtinguts mostra clarament que cada model (Regressió Logística, Random Forest i XGBoost) té fortaleses i debilitats diferenciades, tot i que comparteixen certes

tendències que ens permeten reforçar les nostres hipòtesis inicials. En termes generals, la Regressió Logística ofereix un bon compromís entre sensibilitat i especificitat, amb una AUC de 0.86, evidenciant una bona capacitat per capturar relacions lineals entre les variables. Random Forest, amb una AUC de 0.85, destaca per la seva robustesa i capacitat per manejar relacions no lineals, aconseguint un millor equilibri entre Accuracy i Recall, que és crucial en el nostre context per identificar casos amb pitjor salut mental. XGBoost, amb una AUC de 0.82, també resulta competitiu, però sembla menys consistent en la identificació de casos crítics segons la nostra prioritat (Recall). La comparativa entre aquests mètodes, juntament amb els estudis d'importància de variables i correlacions (heatmap i feature importance), confirma que variables psicològiques com cesd, stai_t i mbi_ex són les més determinants, mentre que sex i glang només apareixen com rellevants en certs contextos per biaixos del dataset i no per relacions reals amb el target. Això reforça la idea que eliminar aquestes variables millora la capacitat dels models per capturar relacions significatives i predir adequadament l'estat mental. A més, en el context de clustering, aquestes millores ajuden a definir estructures més clares, facilitant la classificació binària d'estat mental per proporcionar suport adequat als estudiants segons les seves necessitats. Així, podem concloure que sex i glang reflecteixen més estereotips que utilitat predictiva i la seva exclusió ajuda a construir models més interpretables i pràctics per la nostra finalitat.

Reductors de dimensionalitat

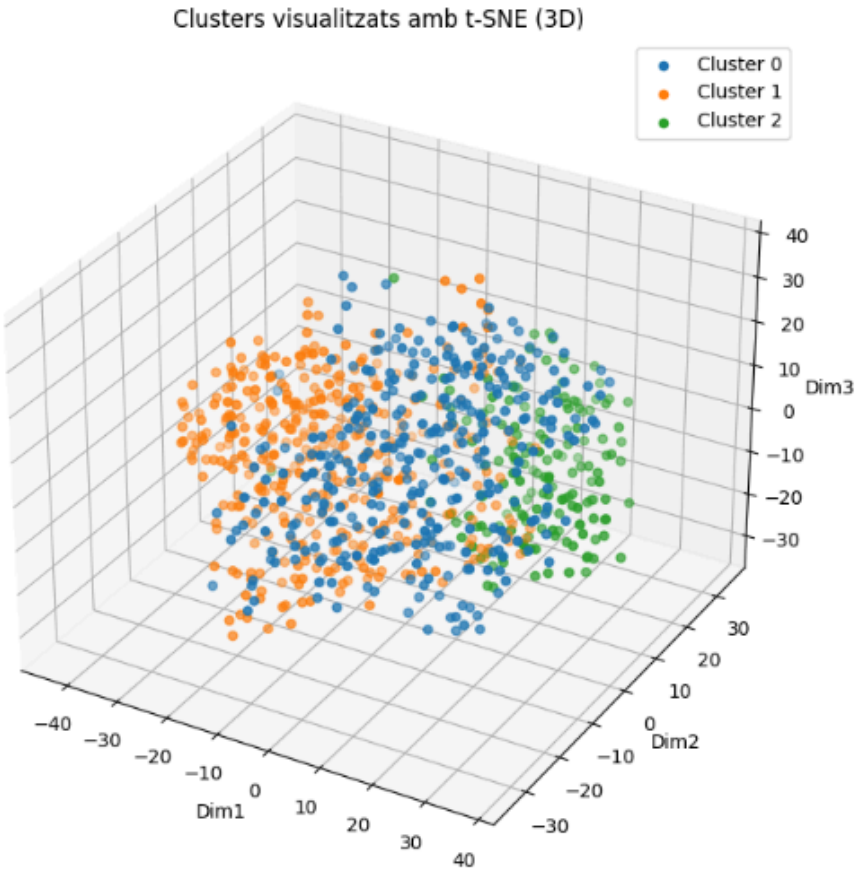
Visualitzar el resultat del clústering és un pas que no és d'especial rellevància, però pot proporcionar una idea intuitiva de com els diferents grups d'estudiants estan distribuïts en funció de les variables analitzades. Per a això, hem comparat dos mètodes de reducció de dimensionalitat: PCA (Principal Component Analysis) i t-SNE (t-Distributed Stochastic Neighbor Embedding), que ens permeten projectar les dades en dues i tres dimensions per tal de visualitzar els resultats posteriorment.

El PCA és un mètode lineal que projecta les dades a un espai de dimensions inferiors, maximitzant la variància explicada per les noves components. És simple i ràpid d'executar, però només funciona bé amb dades lineals.

Per altra banda, el t-SNE s un mètode no lineal que conserva les relacions de veïnatge entre els punts, fent-lo ideal per al nostre cas.

Així doncs, hem fet servir t-SNE per fer les visualitzacions en el nostre projecte.

Cal destacar però, que els clusterings els hem fet amb els dataframes originals per tal de respectar la totalitat de les dimensions i la variabilitat de les dades originals. La reducció de dimensionalitat s'ha fet servir només com a fi de visualització per representar gràficament com s'havien dividit els clústers en el dataframe.



Clusterings i resultats

Per a fer els clústering, vam provar tant algorismes de clustering particionals com jeràrquics, ja que cadascun ofereix avantatges diferents. L'objectiu era observar el comportament de cadascun i comparar els seus resultats per determinar quin d'ells aconsegueix millors divisions.

Per als algorismes particionals, hem escollit utilitzar: kmeans, la seva variació mini-batch kmeans, i gmm (gaussian mixture model). Hem escollit aquests en concret ja que són algorismes senzills, eficients i altament utilitzats que ens permeten assolir el nostre objectiu.

Per als algorismes jeràrquics, únicament hem escollit el algorisme d'agglomerative clustering, ja que ofereix una bona representació sobre com s'agrupen els clústers de manera progressiva, donant una visió clara de com són les relacions entre les dades a diferents nivells.

En tots els casos es busca quina és el nombre més òptim de k clusters per a la clusterització utilitzant mètodes adients per a cada algorisme. En tots els casos hem provat diferents valors de k, en un interval de entre 2 i 8.

El resultat que dóna el clustering són uns labels (etiquetes) que són assignats a cada registre del dataset. Per tal de fer l'assignació, s'afegeix una nova columna al dataframe on se l'hi asigna el número identificador del clúster a cada registre.

	year	stud_h	health	sex_1	sex_2	sex_3	age	...	erec_mean	mbi_ea	part	job	psyt	mbi_cy	Cluster
0	-1.192728	1.929222	-0.733013	1	0	0	-1.328891	...	0.191956	-0.908577	1	0	0	0.636380	0
1	0.508974	-0.332243	0.209584	1	0	0	1.096233	...	-0.317245	0.387023	1	0	0	0.200651	2
2	-0.058260	0.672852	-0.733013	0	1	0	-0.419470	...	-0.317245	-0.260777	0	0	0	-0.670805	0
3	-0.625494	1.615130	1.152182	0	1	0	-0.419470	...	1.210356	-0.692644	0	1	0	-0.017213	0

Aquesta metodologia ens permet, un cop fet el clustering i assignades les id's del cluster a cada registre, agrupar les característiques dels usuaris segons el seu clúster, i calcular mètriques com la mitjana d'una característica (variable) entre tots els usuaris d'un clúster per així estudiar-ne el seu resultat.

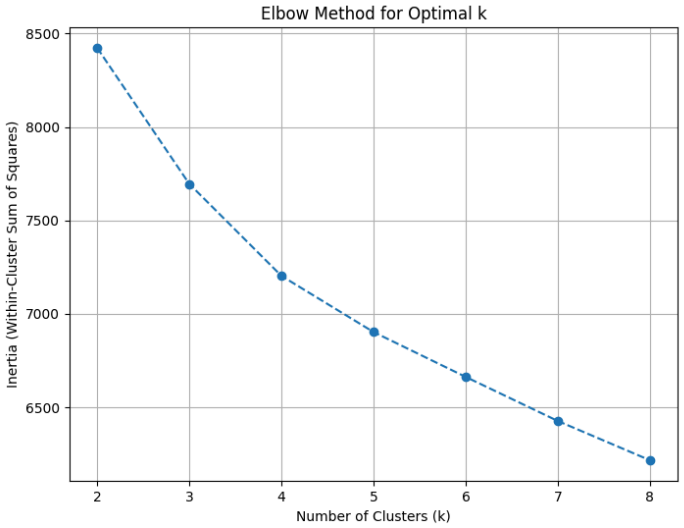
L'últim pas per classificar els clusters d'estudiants segons si tenen millor o pitjor estat de salut mental és fer la mitjana de les variables psicològiques objectiu (nivell d'estrès, ansietat i cansament emocional) en cada clúster. Les tres variables tenen en comú que quan més alt és el seu valor, pitjor estat de salut mental se l'hi atribueix. D'aquesta manera, podem ordenar els clústers tenint en compte la mitjana d'aquestes tres variables i obtenir un llindar mitjançant la mitjana d'aquests tres valors, i així doncs classificar els clústers de tal manera que els que estiguin per sobre del llindar tenen millor estat de salut mental, i els que es troben per sota, un de pitjor.

Kmeans

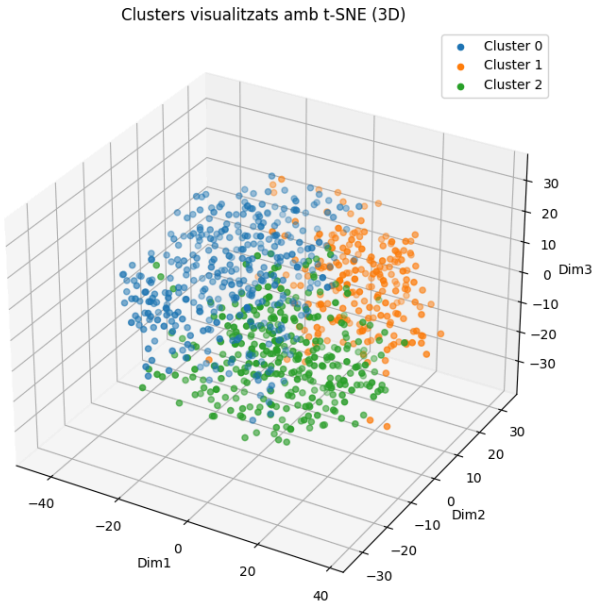
El kmeans agrupa dades assignant-les al clúster més proper mitjançant la distància als centroides. En aquest cas, hem utilitzat dos mètodes per determinar el nombre òptim de k clústers:

El primer mètode és l'elbow method, que consisteix en observar el gràfic que mostra la inèrcia (suma de les distàncies quadrades dins dels clústers) en funció del nombre de clústers k. La idea és trobar el punt on la reducció de la inèrcia comença a ser menys significativa. Aquest punt suggereix el valor òptim de k, ja que afegir més clústers no millora el clustering.

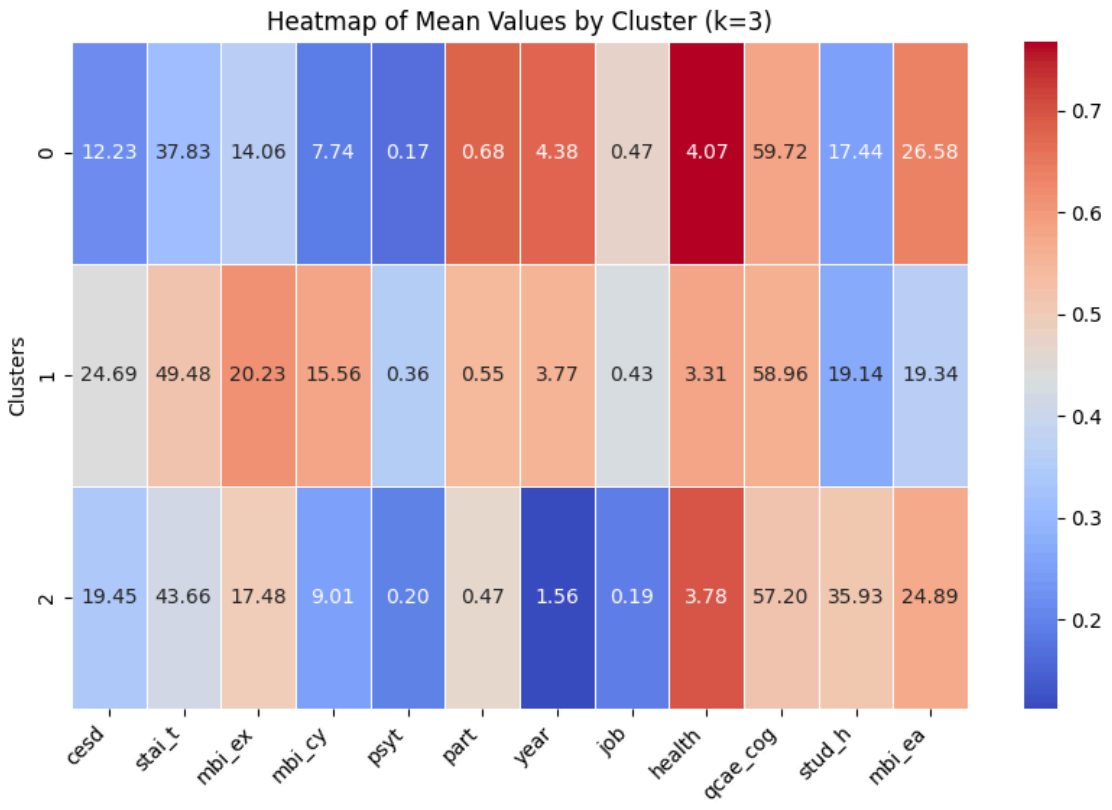
El segon mètode utilitzat és el silhouette score, que mesura com de ben definit està cada clúster, basant-se en la distància mitjana entre punts dins del mateix clúster i la distància mitjana entre punts de clústers diferents. Un silhouette score més proper a 1 indica que els clústers estan ben separats i els punts estan correctament assignats, mentre que un valor proper a 0 indica que els clústers es solapen o els punts estan mal assignats.



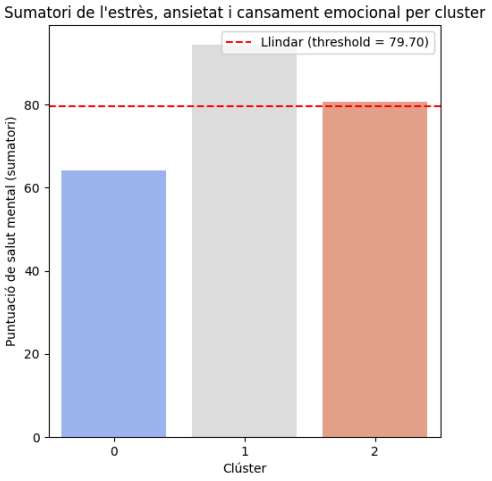
Observant el gràfic es pot veure que hi ha dos punts elbow candidats, en k = 3 i en k = 4. No obstant, el silhouette score resultant dona en aquest cas 3, i per tant, sortim de dubtes i escollim definitivament 3 com el valor més òptim de k clusters.



Per fer la representació visual reduïm les dimensions a 3 utilitzant el reductor de dimensionalidad t-SNE, en aquest cas a 3D.
Es poden veure els 3 clústers diferenciats, però amb punts outliers. És important destacar que els outliers poden aparèixer perquè els clusters s'han generat en el dataframe original sense reduir, on les relacions entre les variables poden ser més complexes. Quan es fa la reducció de dimensionalitat amb t-SNE, aquesta informació es perd en certa mesura, cosa que pot provocar que els punts extremadament llunyans en l'espai original apareguin com a outliers en el gràfic.



En el heatmap podem veure la mitjana de les característiques més rellevants de cada clúster. En el gràfic de la dreta podem veure la classificació i el llindar calculat per a la classificació.



Com es pot veure, el clúster 1 és el que presenta una mitjana superior al llindar, indicant així un pitjor estat de salut mental.
El clúster 2 es troba just en el llindar, indicant així un estat de salut mental estable, sugerint que els usuaris dins d'aquest es troben en un estat moderat d'estrès o amb problemes lleus.
Finalment, el clúster 0 és el que indica tenir el millor estat de salut mental, amb nivells d'ansietat, cansament emocional i estrès més baixos.

Si analitzem en profunditat, el clúster 1 (pitjor estat de salut mental) té les següents característiques que indiquen el perquè del seu estat:

- **Depressió (CESD):** el més alt de tots els clústers.
- **Ansietat (STAI_T):** també el més alt, indicant alts nivells d'estrès o ansietat.
- **Cansament emocional (MBI_EX):** superior als altres grups.
- **Cinisme (MBI_CY):** més alt, potser indicant un despreniment emocional del seu entorn.
- **Salut (HEALTH):** el més baix, indicant que els usuaris són els menys satisfets amb la seva salut.
- **Hores d'estudi (STUD_H):** valor baix, cosa que podria estar associada a dificultats per concentrar-se o manca de motivació.

Per al clúster 0 (estat regular/estable de salut mental):

- **Depressió (CESD):** moderada però menor que el Clúster 1.
- **Ansietat (STAI_T):** també moderada i inferior al Clúster 1.
- **Cansament emocional (MBI_EX):**, relativament baix.
- **Salut (HEALTH):**, valor mitjà-alt, indicant certa satisfacció amb la salut.
- **Hores d'estudi (STUD_H):** el més alt entre els tres clústers, relacionat probablement amb un major compromís acadèmic.
- **Eficàcia acadèmica (MBI_EA):** valor relativament elevat que podria reflectir una bona percepció del rendiment acadèmic.

Per al clúster 2 (millor estat de salut mental):

- **Depressió (CESD):** la més baixa entre els clústers.
- **Parella (PART):** Gran part dels usuaris d'aquest clúster tenen parella sentimental.
- **Visita al psicòleg (PSYT):** valor baix. Degut probablement a que no tenen la necessitat d'anar al psicòleg degut al seu bon estat de salut mental.
- **Cinisme (MBI_CY):** el més baix, suggerint que aquest grup es sent més connectat amb el seu entorn.
- **Salut (HEALTH):**, el més alt, indicant una bona percepció de salut.
- **Eficàcia acadèmica (MBI_EA):** també el més alt, associat a una alta percepció del seu rendiment acadèmic.
- **Hores d'estudi (STUD_H), eficàcia acadèmica (MBI_EA) i nivell cognitiu (QCAE_COG):** en aquest clúster els usuaris són els que menys hores estudien i els que aconsegueixen millors resultats, suggerint que aquest grup té menys problemes per concentrar-se. Altrament, tots tres clústers presenten un nivell cognitiu pràcticament idèntic, indicant que aquest no és d'especial rellevància en els estudis.

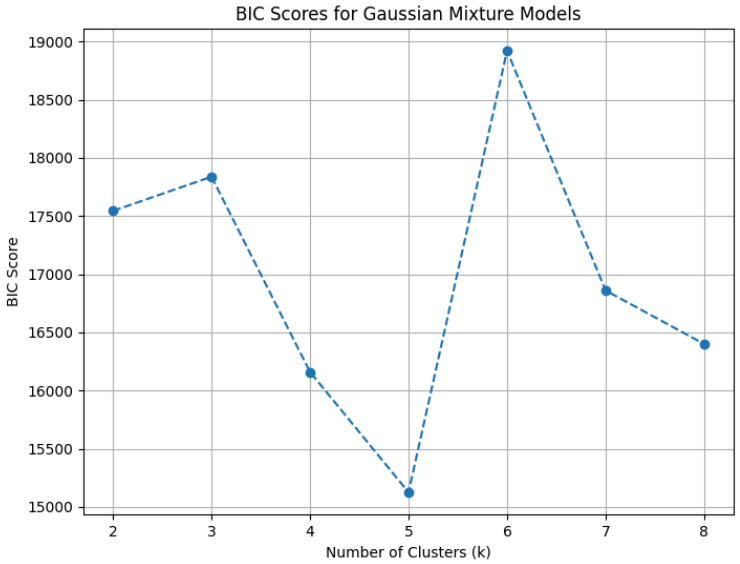
El mini-batch kmeans es tracta d'una versió optimitzada i més ràpida de k-means per a conjunts de dades grans. Com el nostre dataset és relativament petit la diferència de rendiment no és notòria, però el mini-batch-kmeans tarda més temps en completar la seva execució.

En les mateixes condicions, el kmeans tarda 0.4 segons mentre que el mini-batch kmeans tarda 0.63 segons en executar-se, suposant això un empitjorament del 0.63/0.4 x1,575 , degut probablement a que el mini-batch kmeans té una implementació més complexa que pot afegir temps d'execució quan el conjunt de dades és petit.

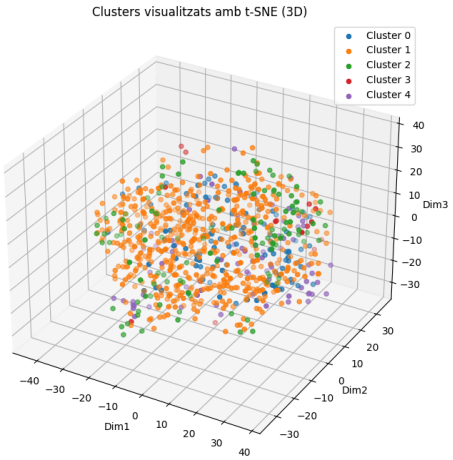
El resultat obtingut és pràcticament idèntic al del kmeans, i per evitar redundàncies descartem analitzar de nou els resultats.

Gmm (Gaussian Mixture Model)

El gmm utilitza models de distribucions gaussianes per agrupar dades en funció de probabilitats i estimació de màxima versemblança. En aquest cas, per trobar la millor k hem fet servir 'model selection', en el qual es fa servir el BIC (Bayesian Information Criterion) per trobar la k més òptima en el gmm. Un bic baix indica bon ajustament entre el model i simplicitat.

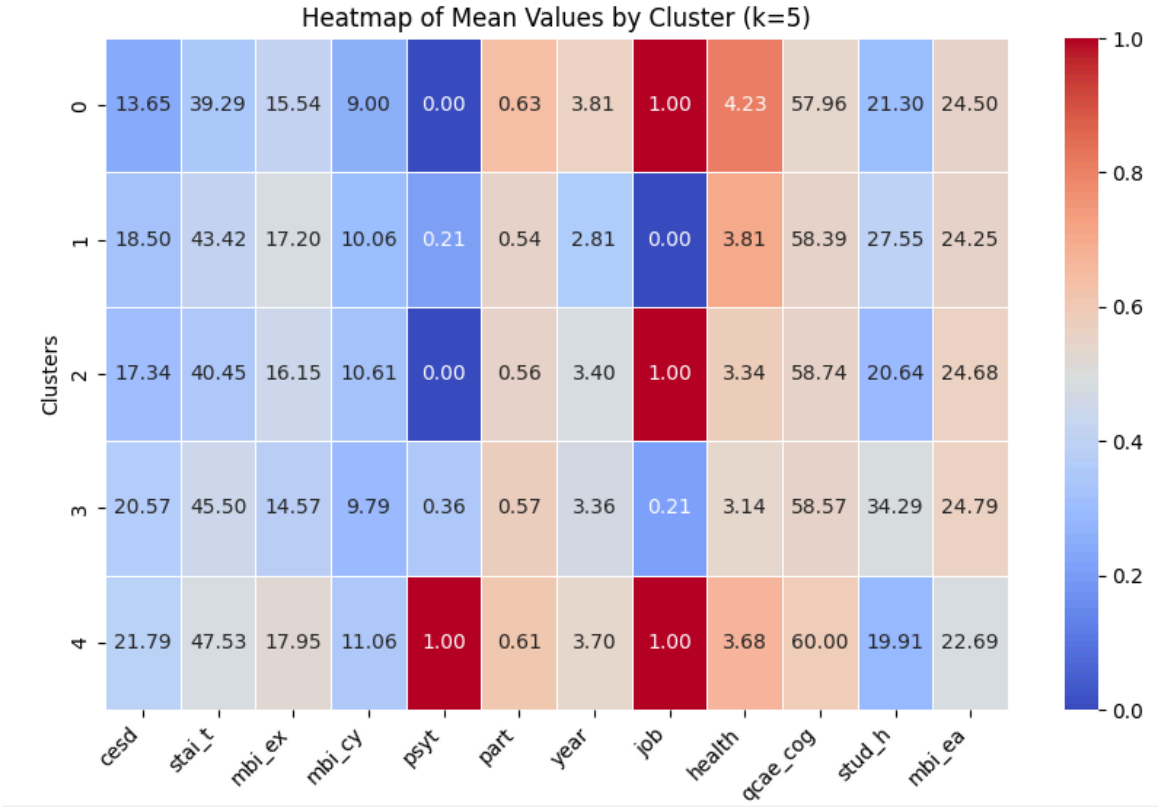


Si observem el gràfic dels BIC scores en cadascun dels clusters, es pot veure que el valor mínim s'atribueix quan k = 5, i per tant, seleccionem aquest valor com al més òptim.



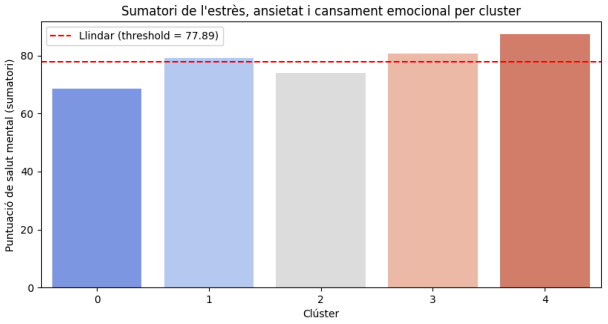
En aquest cas, no es poden apreciar els clústers. Això es deu al fet que el clustering s'ha realitzat amb el dataframe original, sense reduir les variables mitjançant el t-SNE. Quan es fa la reducció de dimensions, com és el cas amb el t-SNE, es perd una part d'informació, la qual cosa fa que els resultats de clustering sobre el dataframe original i el dataframe reduït no siguin exactament iguals. Això provoca que, quan apliquem els labels del clustering sobre el dataframe reduït i els visualitzem, es provoquin distorsions.

No obstant, tot i que la visualització sigui confusa, els resultats del clustering no es veuen afectats.



En aquest cas, la variabilitat en algunes de les característiques entre clusters no és tan definida en comparació a l'algorisme de kmeans degut a que seleccionem un nombre major de k clústers i perquè no hi ha gran variabilitat en el dataset. No obstant, podem comprovar amb el sistema de classificació quins grups es poden classificar en millor o pitjor estat de salut mental.

Pel que fa a la classificació, dels 5 clusters, es pot veure que el 0 és el que presenta millor estat de salut mental i és juntament amb el cluster 4 (pitjor estat de salut mental) els que s'haurien d'estudiar per veure que els fa tindre aquestes característiques. Els altres clusters presenten estats mentals estables, essent lleugerament pitjor el cluster 3 i lleugerament millor el cluster 2.



Si analitzem les característiques del clúster 0 que està per sota el llindar (millorar estat de salut mental):

- **Visita al psicòleg (PSYT):** els usuaris no visiten el psicòleg, probablement degut a que no ho veuen necessari pel seu bon estat de salut mental.
- **Salut (HEALTH):** millor estat de salut física en comparació amb la resta de clústers.
- **Treball i hores d'estudi (JOB, STUD_H):** els usuaris d'aquest clúster tenen un treball estable, cosa que pot indicar estabilitat econòmica el qual pot prevenir estrès o ansietat. No obstant, es veu afectat en les hores d'estudi, es dediquen menys hores, probablement degut a les hores de treball.
- **Parella:** Gran part dels usuaris en aquest clúster (63%) tenen parella sentimental, cosa que pot afavorir l'autoestima, regular nivells d'estrès etc.

Si analitzem les característiques del clúster 4 que està per sobre el llindar (pitjor estat de salut mental):

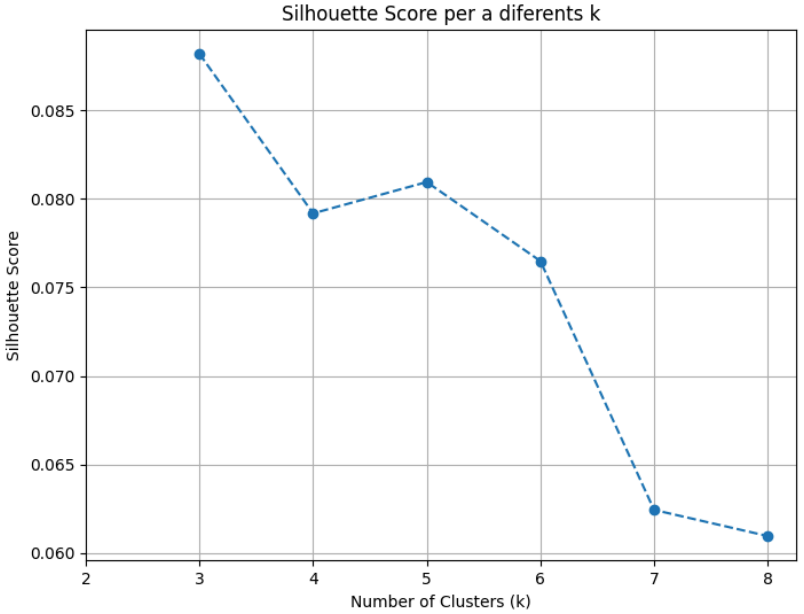
- **Visita al psicòleg (PSYT):** Els usuaris han visitat el psicòleg almenys 1 cop en l'últim any, però segueixen mantenint nivells alts d'estrès, ansietat i cansament emocional, indicant que no es fan prous visites o que les sessions no són completament efectives.
- **Salut (HEALTH):** nivells de salut acceptables, indicant que possiblement no és un factor de risc per als usuaris d'aquest clúster.
- **Eficàcia acadèmica i hores d'estudi (MBI_EA, STUD_H):** nombre més baix de hores dedicades a l'estudi juntament amb el pitjor rendiment acadèmic d'entre tots els clústers, indicant una falta de motivació o dificultats per concentrar-se en els estudis.
- **Treball (JOB):** els usuaris d'aquest clúster treballen, però degut al seu mal estat de salut mental es dedueix que treballen en llocs estressants o mal remunerats, provocant com a conseqüència un mal estat econòmic que pot resultar en un increment de l'estrès i/o l'ansietat.

Si analitzem les característiques dels clústers 1,2 i 3 que tenen mitjanes similars al llindar (estat de salut mental estable):

- **Visita al psicòleg (PSYT):** els usuaris d'aquest grup tenen un índex molt baix de visites al psicòleg, cosa que indica que, tot i tenir un nivell regular de salut mental, no acostumen a buscar ajuda professional.
- **Salut (HEALTH):** els nivells de salut son moderadament bons, però inferiors al clúster amb millor estat de salut mental, cosa que indica que la seva salut física és millorable i que repercuteix a l'estat de salut mental.
- **Eficàcia acadèmica i hores d'estudi (MBI_EA, STUD_H):** el rendiment acadèmic és bo en tot el grup aconseguint resultats similars al clúster 0 amb millor estat de salut mental. No obstant, es podría explicar en el cas dels clústers 1 i 3 degut a que són els que dediquen més hores a l'estudi. El clúster 2 dedica menys hores però aconsegueix uns resultats similars, indicant que aprofita més el temps d'estudi. Es conclou que el seu estat estable de salut mental permet aconseguir bons resultats acadèmics.
- **Treball (JOB):** es pot apreciar que dins el grup, els usuaris del clúster 2 (que tenen millor estat de salut mental) treballen, significant que poden tenir més estabilitat econòmica. En canvi, els usuaris dels altres dos clústers pràcticament no treballen, indicant que probablement tenen un pitjor estat econòmic.

Agglomerative clustering

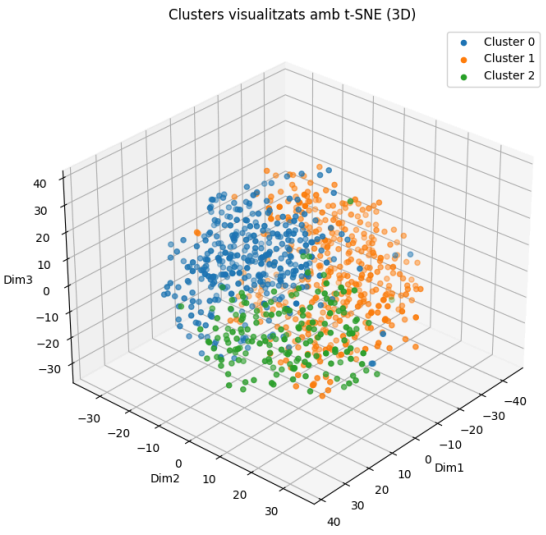
L'Agglomerative clustering és un mètode de clustering jeràrquic on els clústers es construeixen successivament, començant amb cada punt com un clúster individual i fusionant-los gradualment fins a formar un únic clúster global. Per aquest cas, com a mètode per a trobar la millor k (nombre de clusters) hem utilitzat el silhouette score, que mesura com de ben definit està cada clúster.



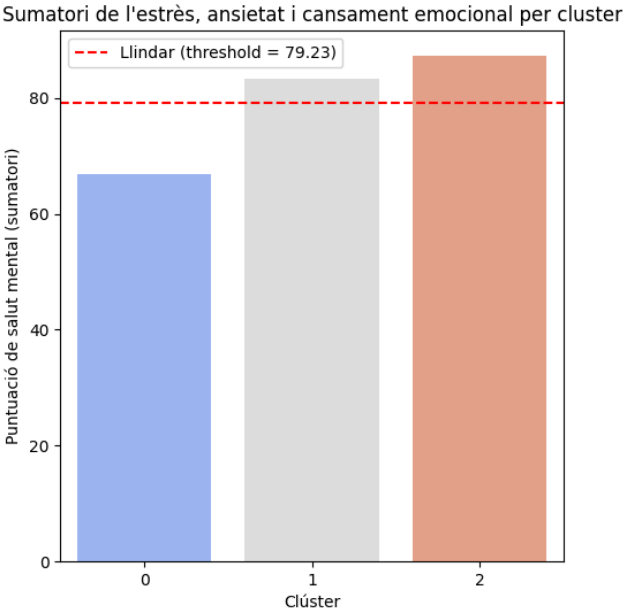
Com es pot observar en el gràfic anterior, el silhouette score més proper a 1 el trobem per per a k = 3, indicant-nos que quan el nombre de clusters es 3, aquests estan ben separats i els punts estan correctament assignats.

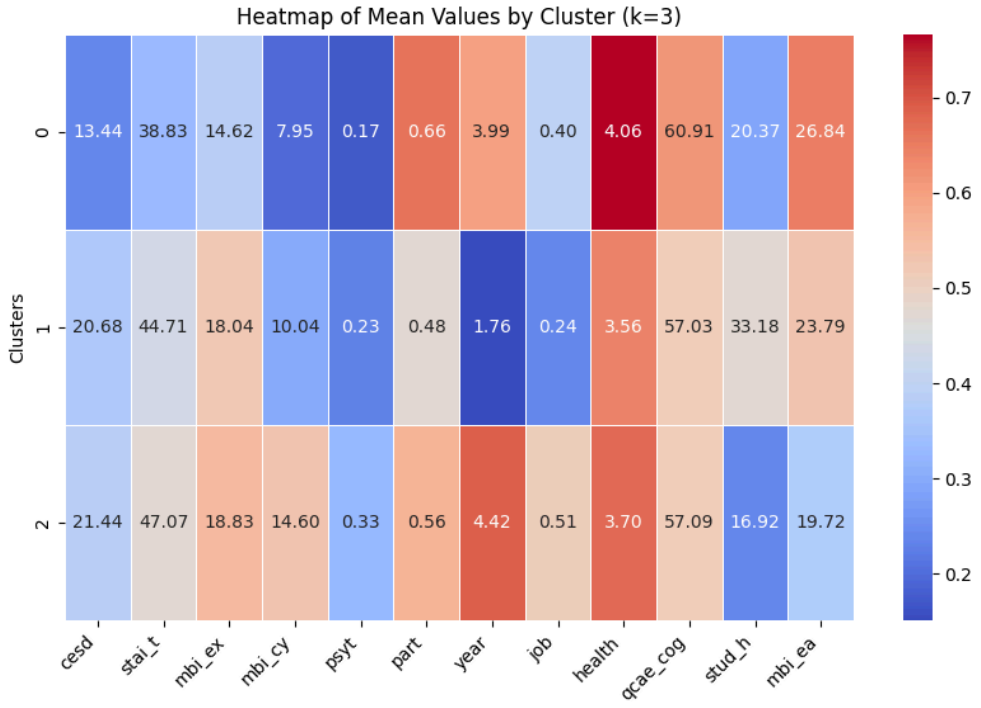
De la mateixa forma que en els casos anteriors, utilitzem t-SNE per a reduir la dimensionalitat fins a 3 dimensions per a poder representar i visualitzar els diferents clusters.

Al Igual que en els altres tipus de clustering, per al cas del clustering aglomeratiu els diferents clusters s'han generat en el dataframe original sense reducció de dimensionalitat de manera que al representar i visualitzar els clústers en l'espai reduït es pot veure cert solapament. Tot i això, es poden identificar 3 grups prou diferenciats.



Per a obtenir una mesura de l'estat de salut mental dels estudiants hem utilitzat el sumatori de les 3 variables objectiu que permeten representar l'estat de salut mental (cesd (ansietat), stai_t (depressió) i mbi_ex (cansament emocional)) i hem establert un llindar igual a la mitjana d'aquest sumatori entre els diferents clusters. D'aquesta manera podem fer un anàlisi més general observant els valors mitjans de l'estat de salut mental per als diversos clusters. S'observa que per al clúster 2, la mitjana de l'estat de salut mental està per sobre del llindar de forma que en aquest grup, els estudiants tenen nivells de depressió, ansietat i cansament emocional més elevats indicant un pitjor estat de salut mental. Pel que fa al clúster 0, representa un grup d'estudiants amb bona salut mental ja que la mitjana del clúster està per sota del llindar. En quant al clúster 1 veiem que la mitjana de l'estat de salut mental està lleugerament per sobre del llindar, de tal manera que ens indica un grup d'estudiants amb un estat de salut mental estable amb nivells depressió i ansietat moderats però alhora un grup al que realitzar un seguiment per tenir controlat.





En l'anterior mapa de calor podem veure el valor mitjà de cada variable per a cada un dels clusters i observar de forma més detallada les característiques de cada grup d'estudiants.

Si observem més detalladament el clúster 2 (estat de salut mental deteriorat) podem veure les següents característiques que indiquen el perquè del seu estat:

- **Depressió (CESD):** nivell de depressió més elevat dels 3 clusters
- **Ansietat (STAI_T):** també el més alt, indicant alts nivells d'estrès o ansietat.
- **Cansament emocional (MBI_EX):** superior als altres clusters
- **Cinisme (MBI_CY):** nivell de cinisme superior que podria indicar un despreniment emocional del seu entorn.
- **Hores d'estudi (STUD_H):** dedicació d'hores als estudis baixa
- **Eficàcia acadèmica (MBI_EA):** percepció del seu rendiment acadèmic com a baix

Per al clúster 1 (estat regular/estable de salut mental):

- **Hores d'estudi (STUD_H):** el més alt entre els tres clústers, relacionat probablement amb un major compromís acadèmic.
- **Visita al psicòleg (PSYT):** valor moderat, alguns dels estudiants d'aquest grup poden ser susceptibles a un deteriorament del seu estat de salut pel que caldria fer un seguiment.
- **Eficàcia acadèmica (MBI_EA):** bona percepció del seu rendiment acadèmic.

Per al clúster 0 (bon estat de salut mental):

- **Depressió (CESD):** nivell de depressió més baix entre els 3 clústers.
- **Ansietat (STAI_T):**nivells d'ansietat considerablement més baixos que ens els grups 1 i 2.
- **Visita al psicòleg (PSYT):** valors baixos de visita al psicòleg entre els estudiants d'aquest grup. Probablement ja que no tenen la necessitat degut al seu bon estat de salut mental.
- **Cinisme (MBI_CY):** nivell de cinisme més baix entre els clústers, els estudiants d'aquest grup es senten més connectats amb el seu entorn.
- **Salut (HEALTH):**, nivells alts que indiquen una bona percepció de la seva salut.
- **Eficàcia acadèmica (MBI_EA):** també el més alt, associat a una alta percepció del seu rendiment acadèmic.

Conclusions

Després de fer l'estudi, hem arribat a la conclusió que 3 és el millor nombre k clusters que podem seleccionar per a divider el conjunt d'estudiants, resultant en una separació més precisa dels diferents estats de salut mental dels estudiants. Dels algorismes provats, tant el Kmeans com l'Agglomerative clustering tenen aquesta k com a nombre més òptim. Alhora, ambdós algorismes donen resultats molt similars, i per tant hem optat per utilitzar kMeans com a mètode principal per la seva simplicitat, escalabilitat i menor cost computacional en comparació amb l'Agglomerative clustering.

Tot i que l'Agglomerative Clustering és útil per entendre millor les relacions jeràrquiques entre els individus, KMeans ofereix una solució més ràpida i pràctica per als nostres objectius.

Després de fer l'estudi i amb els resultats de l'algorisme escollit, ens hem plantejat si hem pogut contestar les preguntes que ens vàrem formular prèviament:

- **Quins grups d'estudiants haurien de rebre major suport per a la seva salut mental?**

Com s'ha pogut veure, la millor manera de separar els estudiants és en 3 grups. Aquests grups es poden classificar segons el seu estat mental en: estat feble, estat regular/estable i en bon estat de salut mental. Per tant, els estudiants que es troben en el grup amb un estat feble de salut mental han de ser la prioritat principal per rebre suport addicional, per ajudar-los amb els seus alts nivells d'ansietat, estrès i cansament emocional.

- **Quines característiques comparteixen els estudiants amb millor/pitjor estat de salut mental?**

En quant als estudiants amb millor estat de salut mental podem destacar el següent:

- Són poc cínics, se senten poc distanciats de la gent i el seu entorn.
- Solen tindre parella sentimental.
- Solen estar en un curs més avançat (mitjans o a finals dels estudis)
- Gaudeixen d'un bon estat de salut.
- Tenen un bon rendiment acadèmic.

És probable que aquests estudiants disposin d'estratègies efectives de gestió de l'estrès i d'un entorn favorable, i és per això que no requereixen d'ajut o suport psicològic en el seu dia a dia.

En quant als estudiants amb pitjor estat de salut mental podem destacar el següent:

- Són més cìnics, se senten més distanciats de la gent i el seu entorn.
- No solen estar molt satisfets amb el seu estat de salut físic.
- Solen tindre un baix rendiment acadèmic
- No solen dedicar suficients hores a l'estudi.
- Reconeixen que tenen un problema psicològic però no solen buscar suficient ajuda.

Aquest grup necessita una atenció especial per abordar els seus alts nivells d'estrès, ansietat i insatisfacció, i per ajudar-los a millorar el seu benestar emocional, físic i acadèmic.

Recomanacions als estudiants desafavorits

L'últim objectiu del projecte era recomanar accions que es poden dur a terme per millorar l'estat de salut mental dels grups en el pitjor estat, basant-nos en el resultat de l'estudi. Així doncs, les nostres recomanacions són les següents:

Intervencions psicològiques:

- Posar en servei dels estudiants programes per gestionar l'ansietat i la depressió, com teràpies individualitzades o grups de suport.
- Augmentar l'accés a serveis psicològics i promocionar-los en l'entorn.

Millora de la satisfacció amb la salut física:

- Promoure hàbits saludables, com l'exercici regular i la bona alimentació.
- Oferir tallers o recursos sobre gestió del son, ja que pot ser un factor important per als estudiants, sobretot en aquest cas, ja que dediquen moltes hores als estudis, sent un posibile indicador de que es dormen poques hores.

Suport acadèmic:

- Tutories per millorar la confiança acadèmica i oferir estratègies per gestionar el temps.
- Espais d'estudi més flexibles i suport emocional en èpoques d'exàmens.
- Tallers o xerrades per promoure tècniques d'estudi més efectives, amb la finalitat de reduir el temps d'estudi general.

Creació d'un entorn social més fort:

- Activitats socials per fomentar connexions entre els estudiants, ajudant a reduir el cinisme.
- Mentoria entre companys per proporcionar suport emocional i acadèmic.

Feina després de la presentació

Després de la presentació es van dur a terme una sèrie de tasques:

- En primer lloc, es va polir el repositori GitHub, acabant els últims directoris, acabant de pulir els scripts etc.

- Es van refer els clusterings amb un dataframe actualitzat, i es va refer l'anàlisi amb els nous resultats.

- Es van revisar i modificar les seccions de feature importance, regressors, etc.