# Projecte IAA

*Esquizofrènia Resistent al Tractament*

**Carlos Palazón Domingo**

# Índex

# 1. Introducció

L'esquizofrènia resistent al tractament (TRS) és un dels principals desafiaments en psiquiatria clínica, ja que afecta entre el 20-30% dels pacients diagnosticats amb esquizofrènia. Aquests pacients mostren una resposta inadequada als antipsicòtics convencionals, mantenint símptomes persistents que deterioren significativament la seva qualitat de vida. Identificar de manera precoç els pacients amb risc de desenvolupar TRS és fonamental per poder ajustar les estratègies terapèutiques i millorar els seus resultats clínics a llarg termini.

L'objectiu d'aquest treball és desenvolupar i comparar diferents models predictius per determinar si un pacient presenta o no TRS. Per entrenar aquests models, hem utilitzat la base de dades `trs_train`, que conté 9000 observacions amb 27 característiques cadascuna, incloent variables demogràfiques, clíniques i genètiques. Un cop identificat el millor model, l'aplicarem per predir 1000 nous casos, amb els resultats a presentar en una competició de Kaggle.

El treball s'estructura en tres etapes clau: primer, un anàlisi exploratori exhaustiu de les dades per comprendre la seva distribució i detectar possibles patrons; segon, el preprocessament necessari per preparar les dades adequadament; i finalment, la construcció i avaluació de tres models predictius diferents, seleccionant aquell que ofereixi el millor rendiment en termes de capacitat predictiva i generalització.

# 2. Anàlisi Exploratori Inicial

Abans d'ajustar qualsevol model predictiu, es realitza una anàlisi exploratòria exhaustiva del conjunt de dades amb l'objectiu d'examinar en detall les variables disponibles, caracteritzar-ne la distribució i identificar possibles valors atípics o inconsistències. Aquest pas permet també determinar les estratègies de preprocessament més adequades, així com les transformacions i codificacions necessàries per garantir la qualitat de les dades abans de la modelització.

## 2.1. Anàlisi estadístic de les variables

A continuació es presenta una anàlisi detallada de certes característiques del dataset, amb l'objectiu de descriure'n les mitjanes, els rangs de variació i les distribucions associades.

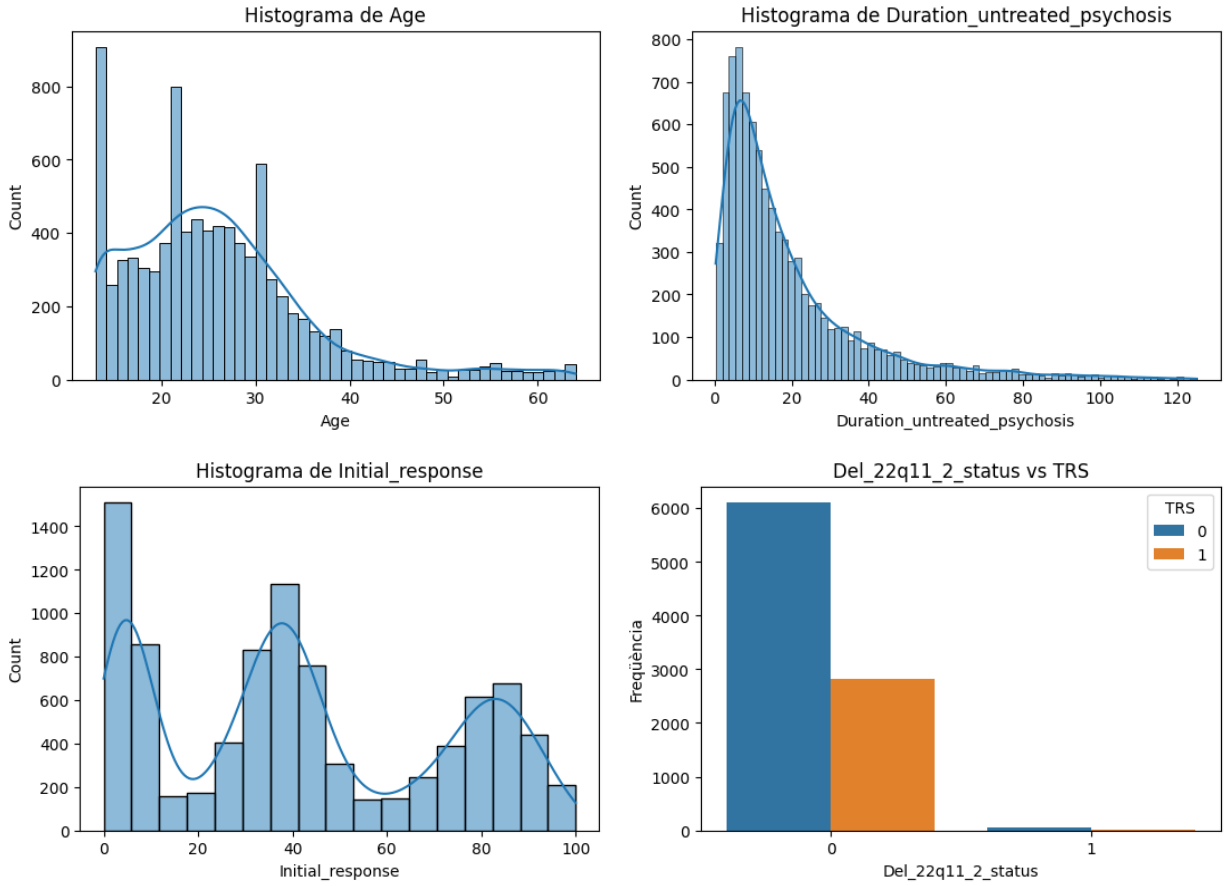



*Figura 1. Distribucions de 4 característiques del dataset (EDA.ipynb).*

Observant l'histograma de l'edat, s'aprecia una distribució clarament asimètrica cap a la dreta, compatible amb una forma similar a una distribució lognormal o gamma. La major part dels valors se situen aproximadament entre els 10 i els 30 anys, amb un màxim de freqüència al voltant dels 20–30 anys. A partir dels 30 anys, la freqüència decreix de manera gairebé monòtona i dona lloc a una cua dreta prolongada que s'estén més enllà dels 60 anys. La variable `Duration_untreated_psychosis` presenta una tendència molt similar, amb una distribució fortament asimètrica cap a la dreta. En aquest cas, però, el pic inicial i el descens posterior són encara més pronunciats, concentrant una proporció molt elevada de casos en valors baixos i generant una cua dreta especialment llarga. Per últim, cal destacar la distribució de la variable `Initial_response`, que presenta un patró clarament multimodal amb diversos pics al voltant de valors baixos, intermedis i alts. Aquest comportament suggereix l'existència de subgrups diferenciats dins la mostra

(per exemple, respostes inicials baixes, moderades i altes), fet que pot tenir implicacions rellevants tant per a la interpretació clínica com per al modelatge posterior. Pel que fa a la resta de variables numèriques, la majoria presenten una distribució aproximadament normal, sense asimetries destacables ni cues extremes remarcables. Aquestes distribucions observades poden suggerir que certes transformacions sobre les variables podrien ser adequades amb l'objectiu d'eliminar les cues pesants i garantir un bon funcionament del model. Això és discutirà més endavant.

Respecte les variables categòriques, podem observar que en alguns casos les classes estàn fortament desbalancejades, com per exemple en `Del_22q11_2_status`:

El gràfic mostra la distribució conjunta de `Del_22q11_2_status` i `TRS` mitjançant un gràfic de barres agrupades, on es veu clarament que la gran majoria de pacients tenen `Del_22q11_2_status = 0`, mentre que el valor 1 és extremadament infreqüent. En concret, tant per als pacients `TRS = 0` com per als `TRS = 1`, gairebé tots es troben a la categoria 0, i només un nombre molt reduït de casos presenta la deleció (1). Aquesta escassetat de casos amb `Del_22q11_2_status = 1` implica que el model disposarà de molt poques observacions per aprendre patrons específics associats a aquesta condició, de manera que les estimacions i l'efecte atribuït a aquesta variable poden ser poc estables i generar prediccions amb alta incertesa per a aquest subgrup minoritari. Aquesta mateixa tendència la segueixen altres variables, com `HLA_A_31_01` o `HLA_B_15_02`. Però, com que el nostre objectiu és predir `TRS`, aquest desbalanceig dels predictors no ens ha de preocupar, ja que afecta únicament la precisió amb què es poden estimar els efectes de les categories més minoritàries, però no desvia de manera significativa l'entrenament global del model ni la seva capacitat de predicció sobre la majoria de pacients.

## 2.2. Balanceig de la classe objectiu

A continuació, s'analitza el balanceig de la classe objectiu per tal de valorar si, en la fase de preprocessament, és necessari aplicar alguna tècnica específica de reequilibri de classes.

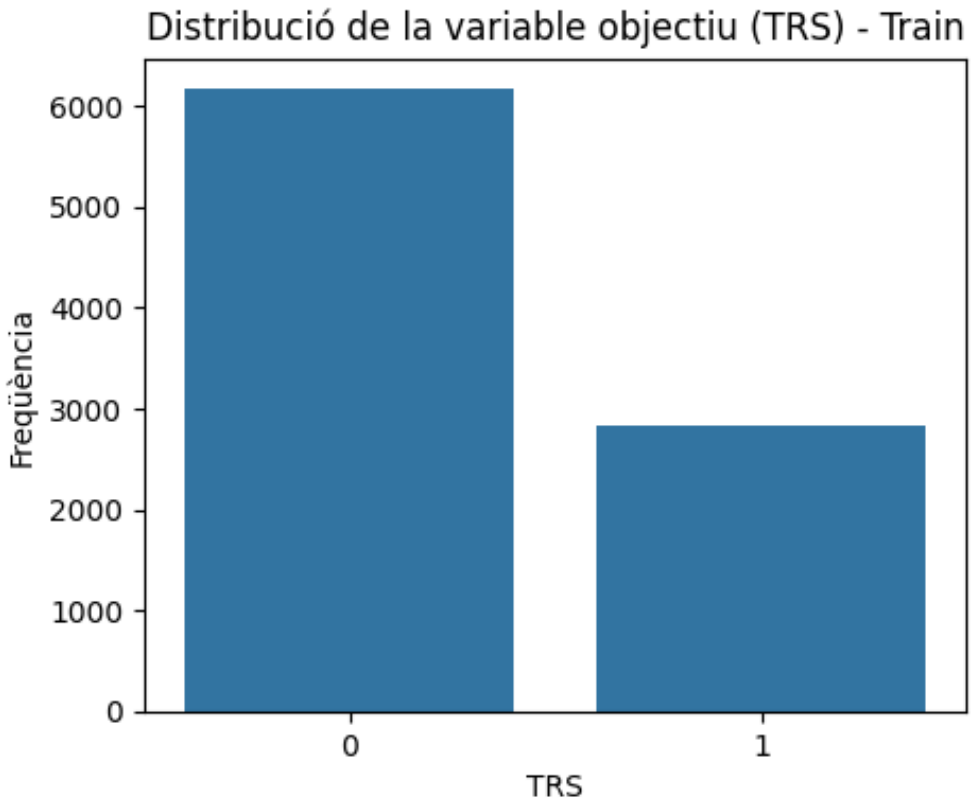


*Figura 2. Gràfic de barres de la freqüència de TRS (EDA.ipynb).*

La distribució de TRS ens mostra una clara proporció desigual entre classes, amb una majoria de casos etiquetats com a 0 i una minoria d'observacions amb etiqueta 1. Aquesta diferència no és extremadament marcada (no es tracta d'un cas de classe rara), però sí suficient perquè un classificador naive que sempre predigués la classe majoritària aconseguís una taxa d'encert raonable sense capturar correctament la classe minoritària. En conseqüència, podem afirmar que **caldrà utilitzar algún mètode de balanceig de classes** per tal de que el model sigui capaç de generalitzar correctament i es garanteixi una detecció més equilibrada d'ambdues etiquetes.

## 2.3. Identificació de missings

En aquesta fase també és fonamental avaluar la completitud del conjunt de dades, ja que la presència de valors mancants pot limitar la informació disponible per a l'entrenament i comprometre el rendiment dels models. Per aquest motiu, es revisarà de manera sistemàtica la presència de missings a les diferents variables, amb l'objectiu de determinar si és necessari aplicar alguna estratègia d'imputació de dades abans de procedir a la modelització. Aquestes són les variables que tenen valors nuls:

```
Duration_untreated_psychosis     128
Lymphocyte_count                1991
Neutrophil_count                1985
Triglycerides                   2453
Glucose                         2619
Alkaline_phosphatase            2938
Polygenic_risk_score               1
IL_17A                             1
```

La majoria de variables es troben completament observades, sense cap valor nul, fet que facilita el seu ús directe en la modelització. En canvi, aquest subconjunt de variables presenta proporcions importants de dades mancants: `Lymphocyte_count` i `Neutrophil_count` acumulen prop de 2.000 valors nuls cadascuna sobre un total aproximat de 9.000 registres, mentre que `Triglycerides` i `Glucose` superen els 2.400 casos amb informació absent. La variable amb més incompleció és `Alkaline_phosphatase`, amb 2.938 valors nuls, i en menor mesura es detecten valors perduts a `Duration_untreated_psychosis`, `Polygenic_risk_score` i `IL_17A`. Aquest patró indica que **serà necessari definir una estratègia específica d'imputació o, eventualment, plantejar-se l'exclusió d'algunes d'aquestes variables en funció de la importància clínica i de l'impacte que pugui tenir aquest volum de dades mancants en el rendiment i la interpretabilitat dels models.**

## 2.4. Outliers

La detecció i gestió d'outliers és un pas clau per garantir la qualitat del conjunt de dades i evitar que valors extrems distorsionin l'ajust dels models. En aquest apartat es descriu el procediment seguit per identificar observacions potencialment anòmales en les principals variables numèriques i es discuteix com aquestes poden afectar tant l'anàlisi exploratòria com el rendiment dels algoritmes de classificació. Per a la detecció

utilitzarem el mètode IQR (interquartile range). Aquest mètode es basa en calcular el rang interquartílic, és a dir, la diferència entre el tercer quartil (Q3) i el primer quartil (Q1), i establir uns límits que defineixen quan una observació pot considerar-se anòmalament alta o baixa.

```
Duration_untreated_psychosis    651
Age                             360
Polygenic_risk_score            125
Ki_associative_striatum          83
Ki_whole_striatum                71
CCL23                            59
TWEAK                            58
IL_17A                           57
BMI                              39
SUVRc_whole_striatum             35
Neutrophil_count                 32
SUVRc_associative_striatum       29
Lymphocyte_count                 26
Alkaline_phosphatase             23
Triglycerides                    22
Glucose                          16
patient_id                        0
Initial_response                  0
```

En el nostre cas, aquest procediment s'ha aplicat a totes les variables numèriques del conjunt de dades, obtenint per a cadascuna el nombre total d'observacions que superen els límits establerts pel mètode IQR. Els resultats mostren que la presència d'outliers no és homogènia entre les diferents variables: mentre que algunes concentren un volum molt elevat de valors extrems, d'altres pràcticament no en presenten. Aquest contrast reflecteix diferències importants en la naturalesa, la variabilitat i la distribució de cada mesura.

La variable que destaca de manera més clara és `Duration_untreated_psychosis`, amb 651 outliers, una xifra molt superior a la de la resta de variables. Aquest resultat és coherent amb el que ja havíem observat en la seva distribució, fortament asimètrica i amb una cua dreta molt marcada. És important remarcar, però, que aquests valors extrems no necessàriament representen errors: poden correspondre a casos clínicament rellevants, de manera que eliminar-los sense una revisió prèvia podria comportar la pèrdua d'informació significativa.

La segona variable amb més outliers és `Age`, amb 360 observacions identificades com a atípiques. L'explicació és similar: l'edat presenta també una distribució amb una cua llarga, cosa que fa que el criteri IQR assenyali com a outliers valors que, en realitat, poden ser perfectament legítims dins del context clínic.

En un segon bloc de variables trobem un nombre moderat d'outliers, com `Polygenic_risk_score`, `Ki_associative_striatum`, `Ki_whole_striatum` i diversos biomarcadors inflamatoris o metabòlics (CCL23, TWEAK, IL-17A, Triglycerides, Glucose, etc.). En aquests casos, els outliers poden reflectir diferències individuals reals o, alternativament, petites imprecisions o soroll en la mesura.

Finalment, la resta de variables presenten un nombre molt reduït d'outliers, fet que indica una distribució més compacta i estable segons el criteri IQR.

Per a considerar el tractar amb els outliets presentats, cal mirar primer si es tracten de valors impossibles. En aquest cas, caldria eliminar els valors directament. Observant el rang de cada variable, veiem que tot i ser outliers, no representen valors impossibles, són observacions vàlides que aporten informació. Per exemple, si observem el rang de la variable `Duration_untreated_psychosis`, veiem que el valor més petit és de 0.3, i el més gran de 125. Tenint en compte que aquesta variable mesura el temps entre l'inici dels símptomes i el començament d'un tractament, en setmanes, podem assegurar que ni 0.3 ni 125 són valors impossibles, són casos reals que cal tenir en compte.

Per això és raonable deixar els outliers, tot i que es podria considerar aplicar alguna tècnica com ara transformacions o winsorització si el model no funciona de la forma esperada.

## 2.5. Anàlisi de correlacions

És important analitzar si hi han variables correlacionades, ja que aquestes no aporten informació adicional i eliminarles podria donar un resultat positiu en el nostre model. Per a estudiar això, observem la matriu de correlacions seguent:

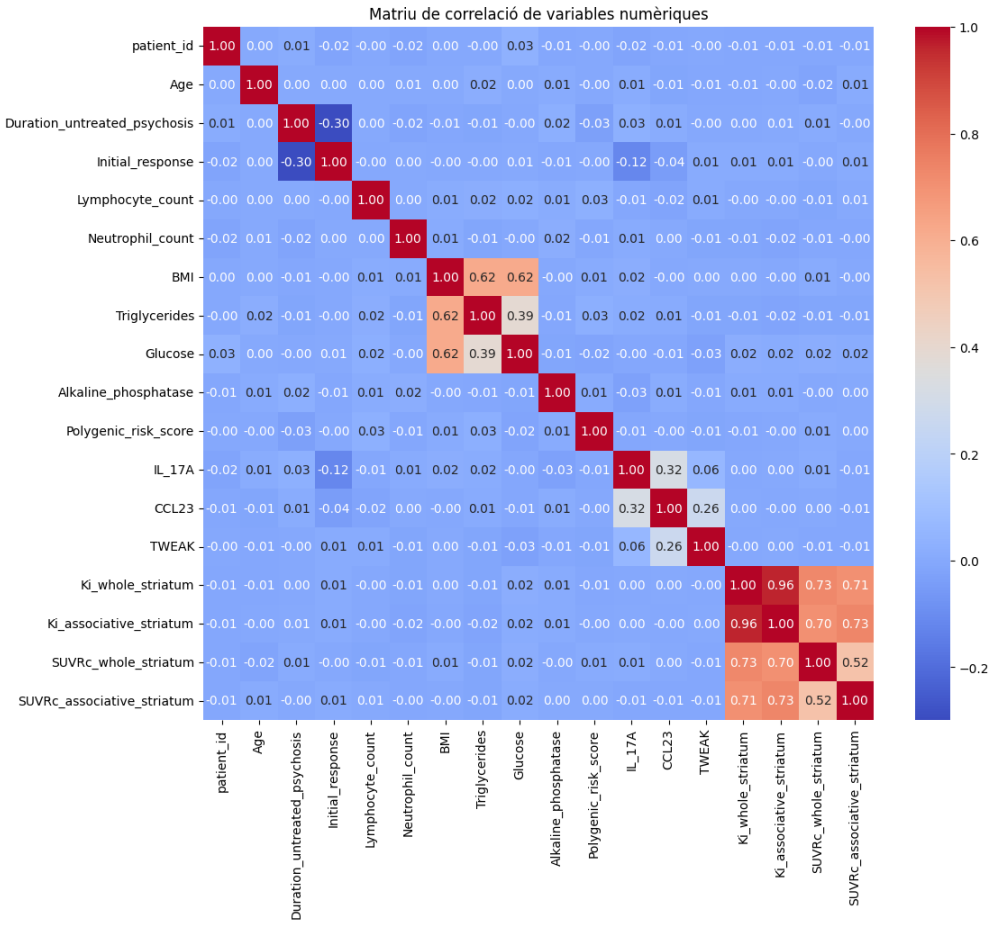



*Figura 3. Matriu de correlacions de les característiques numèriques del dataset (EDA.ipynb).*

Veiem que la majoria de característiques tenen una baixa correlació, a excepció de la part inferior dreta de la matriu, on podem veure varies variables que estan fortament correlacionades. Concretament són: `Ki_whole_striatum`, `Ki_associative_striatum`, `SUVRc_whole_striatum` i `SUVRc_associative_striatum`. Degut a aquesta elevada correlació, és raonable decidir que **cal eliminar**

**com a mínim dues d'aquestes quatre variables per a garantir un millor funcionament del nostre model**. En el preprocessing de cada model s'especifica quines variables s'han el·liminat.

També es mira la correlació de cada una de les característiques numèriques amb la variable resposta:

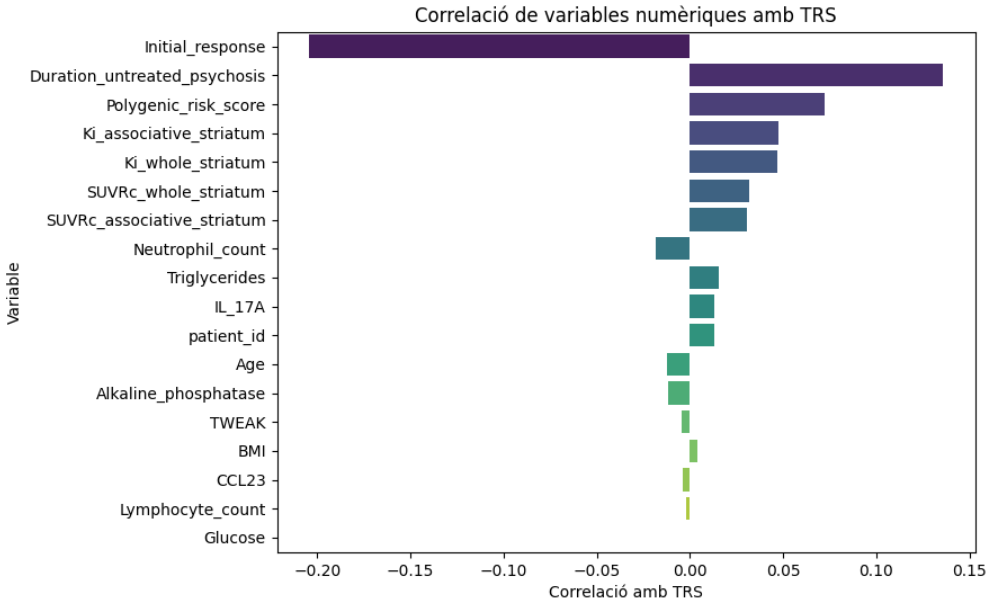



*Figura 4. Correlació de les característiques numèriques amb la variable resposta TRS (EDA.ipynb).*

Veiem que la variable `Glucose` no té correlació amb TRS. Això ens porta a pensar que **eliminarla també és una opció raonable.**

## 2.6. Anàlisi dels components principals

S'ha realitzat una Anàlisi de Components Principals (PCA) sobre el conjunt de dades preprocessat (explicat en detall és endavant), per avaluar la viabilitat de reduir la dimensionalitat de l'espai de característiques. El gràfic de variància explicada acumulada mostra la relació entre el nombre de components principals retinguts i el percentatge d'informació conservada del dataset original.

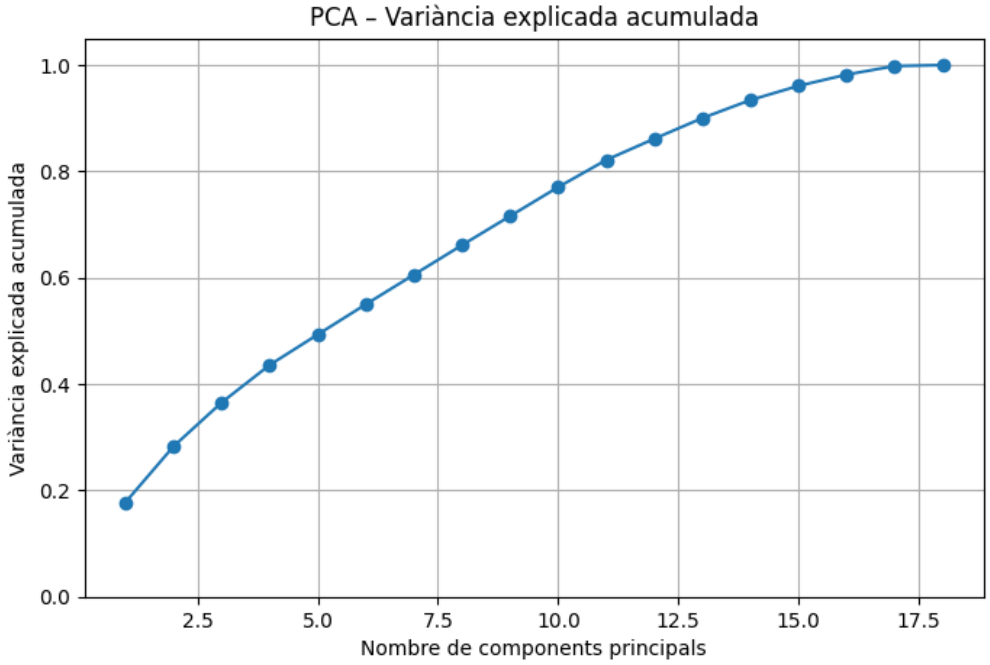



*Figura 5. Variància explicada acumulada per PCA. (EDA.ipynb)*

```
Cumulative variance: [ 0.17751481 0.28281655 0.36524268 0.43587861 0.49318239
0.55001603
 0.60566403 0.6612254  0.71589    0.77035568 0.82141211 0.86136824
 0.90023099 0.93415385 0.96083199 0.98200098 0.99810792 1.        ]
```

En el gràfic, podem observar una absència de colze. La corba presenta un creixement molt suau i quasi lineal, sense el colze característic que sol justificar una reducció de dimensionalitat eficaç. Això indica que la variància està distribuida de manera molt uniforme entre totes les variables, sense factors dominants. Observem també que per conservar un nivell estàndard d'informació del 90%, seria necessari retenir 13 dels 18 components possibles. Això suposa una reducció de dimensionalitat marginal (només 5 dimensions menys), que no compensa la complexitat afegida de la transformació. Per tant podem concloure que **no és necessari realitzar una reducció de dimensionalitat per a les nostres dades.**

# 3. Preprocessament de les dades

Un cop realitzat l'anàlisi exploratori inical, procedim amb el preprocessament de les dades. Cal destacar que no s'ha seguit la mateixa estratègia de preprocessament per a tots els models.

## 3.1. Elimincació de variables

Durant l'anàlisi preliminar, s'ha identificat que algunes variables presenten una correlació elevada entre elles. La presència de variables molt correlacionades pot provocar redundància d'informació i afectar negativament el rendiment i la interpretabilitat dels models predictius. Per tal de minimitzar aquest efecte, s'han eliminat les següents variables: `Ki_associative_striatum` i `SUVRc_associative_striatum`. Aquestes variables estan fortament correlacionades entre si i amb altres mesures relacionades amb el metabolisme cerebral, per la qual cosa la seva exclusió ajuda a reduir la multicol·linearitat sense perdre informació rellevant.

A més, s'han descartat dues variables addicionals per raons específiques: `patient_id`, un identificador únic de cada participant i, per tant, no aporta informació explicativa sobre la variable objectiu. `Glucose`, ja que les anàlisis inicials no mostren cap relació significativa amb la variable objectiu `TRS`, de manera que la seva inclusió podria introduir soroll sense benefici predictiu. L'eliminació d'aquestes variables té com a objectiu optimitzar la qualitat de les dades i millorar la robustesa i interpretabilitat dels models de predicció.

## 3.2. Codificació de variables categòriques

Per tal de poder utilitzar les variables categòriques en els nostres models predictius, cal transformar-les a un format numèric. Tant els models basats en distàncies, com ara SVM o regressió logística, com els models basats en arbres, com XGBoost, no poden processar directament variables codificades com a text o caràcters. Per aquesta raó, s'ha aplicat una codificació numèrica a totes les variables categòriques. En concret, s'ha utilitzat el mètode One-Hot Encoding, que crea una columna binària per a cada categoria de la variable, indicant la presència (1) o absència (0) de la categoria en cada observació. Aquest procés s'ha implementat amb la funció `OneHotEncoder` de la llibreria pandas, que facilita la generació automàtica de totes les columnes necessàries per a cada variable categòrica. Aquesta codificació garanteix que les variables siguin compatibles amb tots els models seleccionats, mantenint al mateix temps la informació original de les categories sense introduir ordres o jerarquies artificials.

## 3.3. Imputació

Com s'ha observat durant l'anàlisi inicial, algunes variables presenten un nombre considerable de valors faltants. Per tal de poder utilitzar aquestes característiques en els models predictius, és necessari substituir aquests valors amb estimacions raonables, procés conegut com a imputació.

Per a les variables numèriques, s'ha utilitzat la funció `SimpleImputer` per imputar utilitzant la mediana. Aquesta estratègia d'imputació s'ha escollit per diverses raons:

- **Robustesa davant outliers**: La mediana no es veu afectada per valors extrems, a diferència de la mitjana. Com s'ha observat en l'anàlisi exploratori, variables com `Duration_untreated_psychosis` i `Age` presenten outliers significatius. En aquests casos, la mitjana podria distorsionar la imputació cap a valors no representatius de la tendència central real de les dades.

- **Adequació per a distribucions asimètriques**: Moltes de les variables numèriques del dataset (com `Duration_untreated_psychosis`, `Lymphocyte_count`, `Triglycerides`) presenten distribucions clarament no normals amb asimetria cap a la dreta. En aquests casos, la mediana proporciona una millor representació del valor "típic" que la mitjana aritmètica.
- **Preservació de la distribució original**: L'ús de la mediana garanteix que els valors imputats es mantinguin dins del rang central de les dades observades, evitant introduir bias sistemàtic que podria afectar el rendiment dels models predictius.

Per a les variables categòriques, s'ha utilitzat la mateixa funció, pero en aquest cas imputant el valor més frequent. Aquesta decisió s'ha pres per les següents raons:

- **Naturalesa discreta de les variables**: A diferència de les variables numèriques, les variables categòriques (com `Sex`, `Del_22q11_2_status`, `HLA_A_31_01`, `HLA_B_15_02`) tenen un nombre finit de categories predefinides. En aquests casos, la mediana o mitjana no tenen sentit interpretatiu, i només és vàlid imputar amb una de les categories existents.
- **Preservació de la distribució majoritària**: Utilitzar el valor més frequent (moda) manté la proporció de les categories originals i minimitza la introducció de bias en les variables categòriques.
- **Simplicitat i interpretabilitat**: La imputació per moda és una estratègia conservadora i fàcilment interpretable que no requereix suposicions complexes sobre la distribució de les dades categòriques. Això garanteix que els valors imputats siguin coherents amb els patrons observats en les dades d'entrenament.

Aquest procés assegura que totes les observacions puguin ser utilitzades en els models predictius sense perdre informació ni introduir biaixos significatius derivats de l'imputació.

## 3.4. Normalització

Els models basats en distàncies, com el SVM i la regressió logística, requereixen normalització de les dades per garantir un funcionament òptim. Aquesta necessitat es deu al fet que aquests algorismes són sensibles a l'escala de les variables. El SVM calcula distàncies euclidianes entre punts en l'espai de característiques, de manera que si una variable té un rang molt més ampli que les altres, dominarà el càlcul de distàncies i el model prioritzarà aquesta característica independentment de la seva rellevància predictiva real.

A més, la regressió logística utilitza algoritmes d'optimització com el gradient descent per minimitzar la funció de pèrdua. Quan les variables tenen escales molt diferents, el gradient esdevé irregular i la convergència és més lenta o pot fins i tot fallar. L'escalat garanteix que totes les variables contribueixin de manera equilibrada al procés d'optimització.

Per aquests motius, s'ha aplicat `StandardScaler` de scikit-learn al SVM i a la regressió logística. Aquesta transformació estandarditza cada variable per tenir mitjana 0 i desviació estàndard 1, assegurant que totes les característiques tinguin la mateixa importància inicial en el model. Cal destacar que XGBoost, basat en arbres de decisió, no requereix aquest pas ja que les divisions dels arbres es basen en llindars relatius dins de cada variable, no en distàncies absolutes entre observacions.

## 3.5. Balanceig de classes

Tal com s'ha observat en l'anàlisi exploratori, la classe objectiu presenta un fort desbalanceig, amb molts més casos de no-TRS que de TRS. Per tal de corregir aquest desbalanceig i evitar que els models es biaixin cap a la classe majoritària, s'ha aplicat ponderació de classes, de manera que els errors en la predicció de la classe

minoritària tenen més pes durant l'entrenament. En els models implementats amb scikit-learn, aquesta ponderació s'ha aplicat mitjançant l'argument `class_weight='balanced'`. En la regressió logística personalitzada, els pesos de classe s'han implementat de forma manual dins de la pròpia classe (explicat en detall més endavant). Aquesta estratègia assegura que els models donin més importància a la correcta predicció dels pacients TRS, millorant la sensibilitat i F1-score de la classe minoritària sense perdre precisió en la classe majoritària.

## 3.6. Split del dataset

Per poder saber si els nostres models funcionen bé amb dades noves que no han vist mai, hem de dividir el dataset original en dos conjunts separats: un conjunt d'entrenament, que és el que farem servir per entrenar els models, i un conjunt de validació (o test), que ens servirà per comprovar com de bé es comporten els models quan han de fer prediccions sobre dades que no han utilitzat durant l'entrenament. És molt important fer aquesta divisió abans d'aplicar cap tipus de preprocessament (com la imputació de valors perduts, la normalització o la codificació de variables categòriques). Si féssim el preprocessament abans de la divisió, estaríem deixant que informació del conjunt de validació "es filtrés" cap al conjunt d'entrenament. Això faria que el model semblés funcionar millor del que realment ho fa, perquè ja hauria vist indirectament part de la informació de les dades de validació. S'ha decidit fer la divisió de forma que el 80% de les dades vagin a entrenament i el 20% a validació. Aquesta proporció és bastant estàndard i ens dóna un bon equilibri: amb el 80% (que són 7200 mostres del total de 9000) tenim prou dades perquè el model pugui aprendre bé els patrons, mentre que amb el 20% restant (1800 mostres) disposem d'un conjunt prou gran per avaluar de manera fiable com generalitza el model. Aquesta divisió ens permet entrenar models robustos sense perdre capacitat d'avaluació.

# 4. Support Vector Machine

El primer model que s'ajustarà és el Support Vector Machine (SVM).

## 4.1. Entrenament del model

En el context de la predicció de l'esquizofrènia resistent al tractament (TRS), s'ha prioritzat l'ús del F1-macro com a mètrica d'avaluació, ja que permet equilibrar el rendiment entre ambdues classes (TRS i no-TRS) independentment del desequilibri del dataset. Aquesta mètrica calcula l'F1-score per a cada classe per separat i després en fa la mitjana aritmètica sense ponderar per la freqüència de les classes, garantint que el model no ignori la classe minoritària.

Per a entrenar el model SVM, s'ha utilitzat un procediment de grid search amb validació creuada de 5 particions (5-fold CV) amb l'objectiu de trobar la combinació d'hiperparàmetres que dona millor rendiment sobre les dades d'entrenament.

Els hiperparàmetres que s'han provat es mostren a la taula següent, i els que estàn en negreta han sigut els escollits:

| **Hiperparàmetre** | **Valors provats** |
|---|---|
| `C` | 0.01, 0.1, 1, 10, 100 |
| `gamma` | 0.001, 0.01, 0.1, 1 |
| `kernel` | rbf, linear |

El paràmetre C controla el compromís entre maximitzar el marge i minimitzar els errors de classificació. Valors petits permeten un marge més ampli a costa d'acceptar més errors, mentre que valors més grans penalitzen més els erros i tendeixen a ajustar-se més a les dades, amb un risc més elevat d'overfitting. En el nostre cas s'han considerat els valors `C = [0.01, 0.1, 1, 10, 100]`, que formen una escala logarítmica i cobreixen des de una regularització forta fins a una regularització molt laxa. Aquest rang permet testar models des de molt simples (amb alta capacitat de generalització però possiblement inframodelats) fins a models molt flexibles que poden capturar patrons més complexos.

Per altra banda, el tipus de kernel determina la forma de la funció de similaritat utilitzada per el SVM i, per tant, el tipus de frontera de decisió que el model pot apendre. El kernel lineal correspon a un hiperplà lineal en l'espai original de característiques, mentre que el kernel RBF permet modelar fronteres no lineals. S'han seleccionat aquests dos kernels per dues raons principals. D'una banda, el kernel lineal és adequat quan la relació entre les característiques i la resposta és aproximadament lineal o quan el nombre de característiques és molt alt, oferint un model més parsimoniós. D'altra banda, el kernel RBF actua com a opció flexible per defecte quan es volen capturar patrons no lineals sense especificar manualment transformacions de les variables. Incloure ambdós kernels en la graella permet que el procés de validació creuada determini empíricament quin enfocament (lineal o no lineal) és més adequat per al conjunt de dades analitzat.

Per últim, tenim el paràmetre gamma, que controla la influència de cada mostra d'entrenament quan s'utilitza un kernel de tipus radial (RBF). Valors baixos de gamma fan que l'efecte de cada punt s'estengui sobre una

regió més àmplia de l'espai de característiques, generant fronteres de decisió més suaus i generalitzades; en canvi, valors alts de gamma fan que la influència de cada punt sigui molt local, produint fronteres més irregulars i potencialment sobreajustades. Per tal de capturar aquests comportaments, s'ha definit el rang `gamma = [0.001, 0.01, 0.1, 1]`. Aquest conjunt de valors, també en escala logarítmica, permet explorar des de configuracions molt suaus fins a configuracions molt detallades de la frontera de decisió, i és adequat quan es desconeix a priori el grau de no linealitat present en les dades.

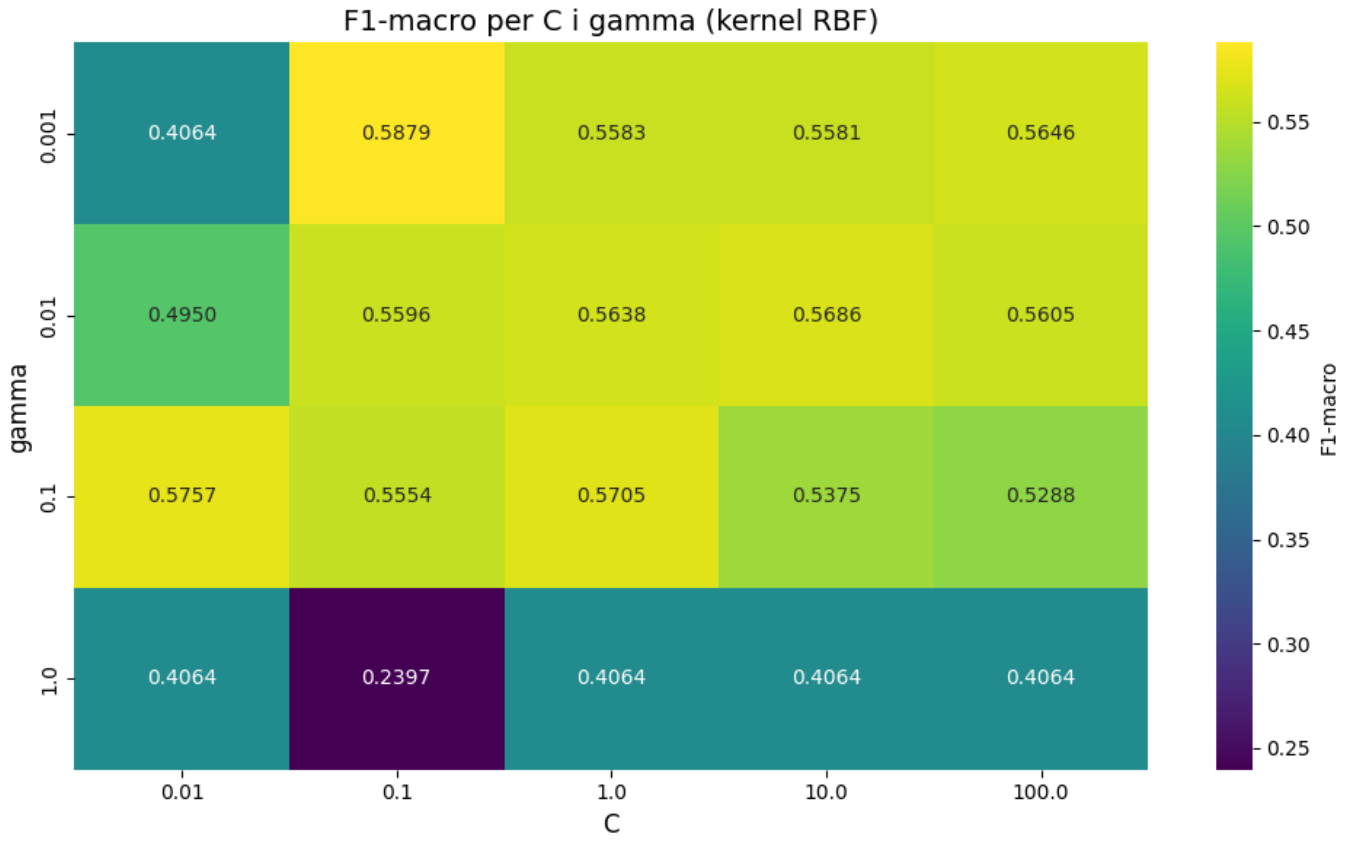



*Figura 6. Heatmap dels hiperparàmetres C i gamma (SVM.ipynb).*

En el heatmap podem observar el rendiment del model SVM, mesurat amb F1-macro, per a cada combinació de C i gamma, amb el kernel RBF. Els resultats mostren un patró clar: `gamma = 1.0` produeix el pitjor rendiment global, amb un F1-macro mínim de 0.2397 quan `C = 0.1`, indicant sobreajustament sever. L'àrea fosca (morada) en aquesta fila del heatmap reflecteix que el model memoritza el conjunt d'entrenament sense capacitat de generalització. Aquest valor tan alt de gamma fa que cada punt d'entrenament tingui un radi d'influència molt petit, creant una frontera de decisió extremadament complexa i irregular que s'ajusta massa als detalls específics del train set. Al reduir el valor de gamma a 0.1, el rendiment del model millora substancialment, obtenint valors de F1-macro entre 0.5288 i 0.5757, representats amb colors grocs i verds clars al heatmap. Aquesta millora es deu al fet que, amb un gamma més baix, cada punt d'entrenament influeix en una àrea més àmplia, permetent que el model construeixi una frontera de decisió més suau i generalitzable. Amb `gamma = 0.1`, observem que el rendiment és relativament estable per a tots els valors de C, amb una lleugera millora quan C pren valors intermedis i alts. Continuant amb la reducció de gamma a 0.01, el model manté un bon rendiment amb valors de F1-macro entre 0.4950 i 0.5686, representats amb tonalitats grogues i verdoses. Aquesta zona intermèdia mostra que el model encara és capaç de capturar els

patrons rellevants de les dades sense caure en sobreajustament. És interessant observar que amb aquest valor de gamma, els valors més alts de C (de 0.1 a 100) tendeixen a produir millors resultats, suggerint que el model es beneficia d'una penalització més forta dels errors de classificació quan la frontera de decisió és més suau. Finalment, amb `gamma = 0.001`, el rendiment varia considerablement depenent del valor de C. Per valors baixos de C (0.01), obtenim F1-macro al voltant de 0.4064, mentre que per C més alt (0.1), el rendiment millora fins a 0.5879, la qual cosa representa un dels millors resultats observats al heatmap. Tanmateix, amb valors de C encara més alts (1, 10, 100), el rendiment es manté estable al voltant de 0.5581-0.5646. Això suggereix que amb un gamma molt baix, el paràmetre C juga un paper més crític per aconseguir un bon equilibri entre la simplicitat del model i la seva capacitat predictiva.

## 4.2. Avaluació del model

Un cop tenim el model amb els millors hiperparàmetres segons la validació creuada, procedim a la seva avaluació, per analitzar les seves mètriques i el seu funcionament amb dades noves.

Comencem analitzant la matriu de confussió i la corba ROC:

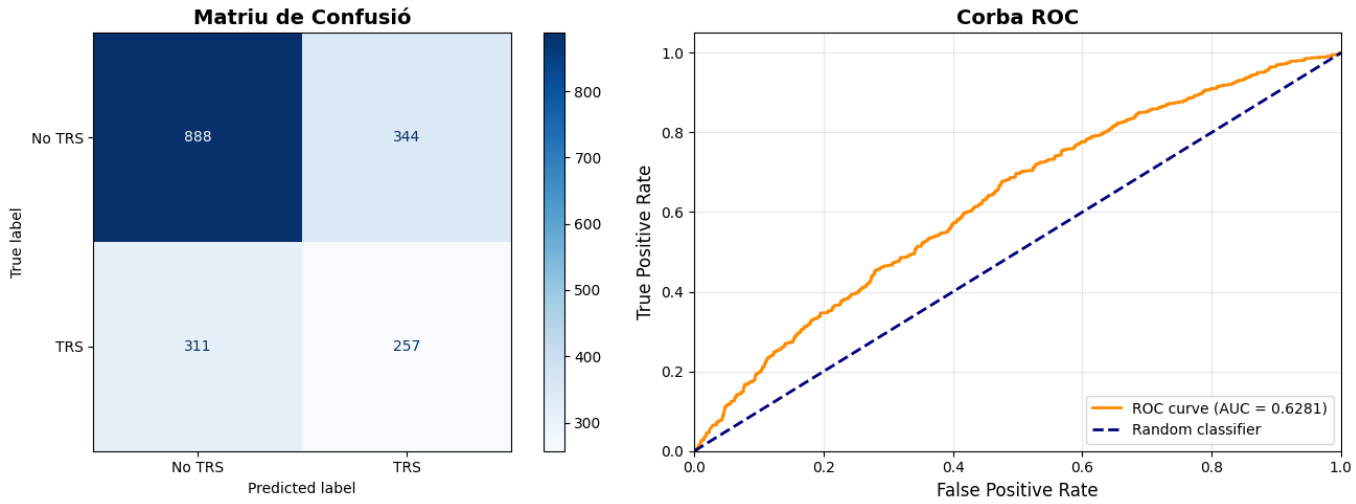



*Figura 7. Matriu de confusió i corba ROC del SVM (SVM.ipynb).*

La matriu de confusió revela un desequilibri significatiu en el rendiment del model entre les dues classes. Per a la classe No-TRS (majoritària), el model aconsegueix 888 veritables negatius però genera 344 falsos positius. Per a la classe TRS (minoritària), s'observen 257 veritables positius però 311 falsos negatius. En general, el model no té un molt bon funcionament, ja que falla moltes de les prediccions, i això en termes pràctics pot suposar un problema. Depenent de l'objectiu del model, es podria plantejar el canviar la mètrica a optimitzar em l'entrenament. Per exemple, veiem que el nostre model té 311 falsos negatius. Si la nostra prioritat és detectar el màxim nombre de pacients amb TRS, aquí estem tenint masses errades de predicció, i canviar la mètrica a F1 seria una bona opció.

Analitzant també la corba ROC, veiem que mostra un AUC de 0.6281, que indica una capacitat de discriminació limitada del model. Aquest valor, només lleugerament superior al 0.5 d'un classificador aleatori, suggereix que el model té dificultats per separar efectivament les dues classes. La forma de la corba, que s'apropa a la línia diagonal en diversos punts, confirma aquesta limitació en la capacitat predictiva.

Ara, analitzem les mètriques del nostre model tant sobre el conjunt d'entrenament com en el conjunt de validació, per tal de veure com funciona per a cada cas i si pateix d'overfitting o underfitting:

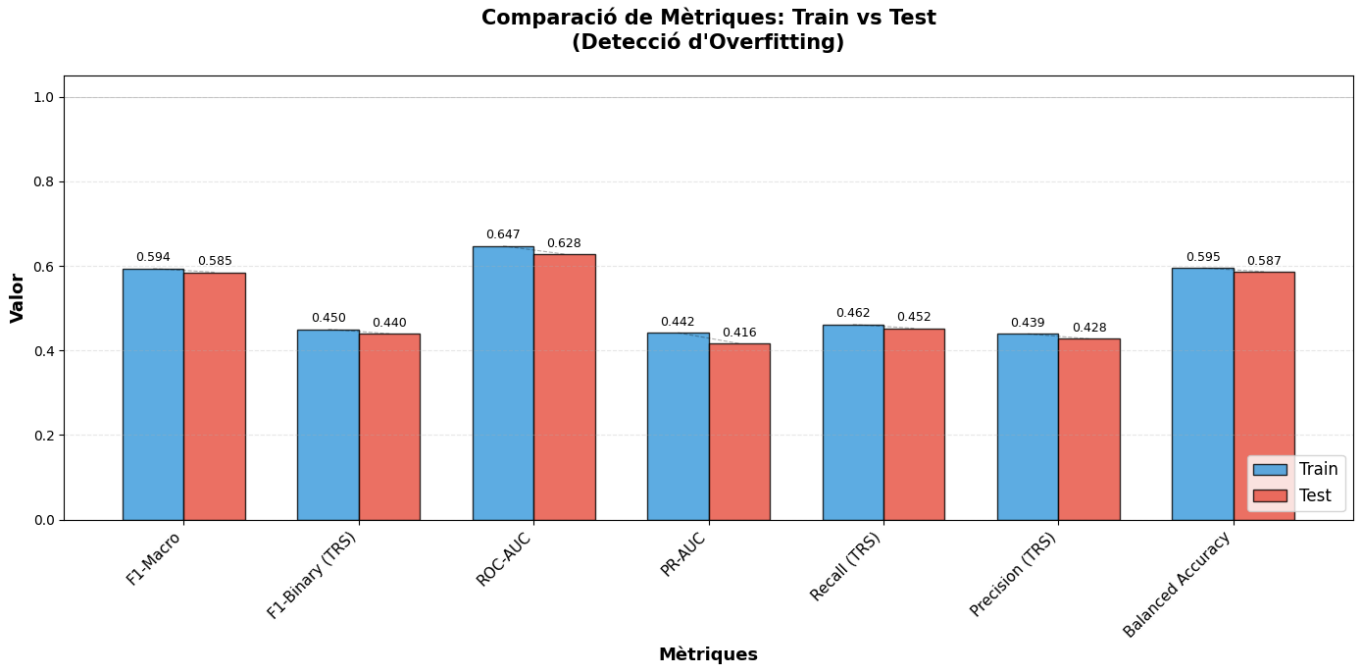



*Figura 8. Gràfics de barres de les mètriques del SVM (SVM.ipynb).*

La comparació entre les mètriques de train i test indica que el model SVM no presenta overfitting rellevant. Tot i que els valors en el conjunt d'entrenament són lleugerament superiors als del conjunt de test, les diferències observades són molt petites i es mantenen estables en totes les mètriques. La diferència més gran es dona en el ROC-AUC, que passa de 0.647 en train a 0.628 en test, amb una diferència de només 0.019 punts. Aquesta variació és mínima i entra dins del que és esperable quan el model s'aplica a dades no vistes, per la qual cosa no indica una pèrdua significativa de capacitat de generalització. Pel que fa a les mètriques específiques de la classe TRS, també veiem diferències molt petites. El F1-Binary (TRS) baixa només 0.010 punts i el Recall (TRS) cau també 0.010 entre train i test. Això vol dir que el model manté aproximadament la mateixa capacitat per identificar pacients amb TRS tant en les dades d'entrenament com en les noves. Tanmateix, els valors absoluts són bastant modestos (al voltant de 0.44-0.45), la qual cosa indica que el model falla en detectar més de la meitat dels casos reals de TRS, cosa que pot ser problemàtica en un context clínic. El F1-Macro, que és la mètrica que hem utilitzat per optimitzar els hiperparàmetres, es manté molt estable amb una diferència de només 0.009 punts entre train (0.594) i test (0.585). Això és un bon senyal perquè indica que el rendiment és equilibrat entre ambdues classes. De manera similar, la Balanced Accuracy només varia 0.008 punts, confirmant que l'equilibri entre la capacitat de detectar casos positius (sensibilitat) i negatius (especificitat) es manté quan apliquem el model a dades noves.

En resum, totes les diferències entre train i test són inferiors al 3%, cosa que és totalment normal i no indica cap problema de sobreajustament. El model generalitza correctament, però el seu rendiment global és moderat: té dificultats per discriminar bé entre les dues classes i especialment per detectar els casos de TRS, com mostren els valors relativament baixos de ROC-AUC i Recall.

# 5. XGBoost

Construïm ara un model XGBoost. A diferència del Support Vector Machine i de la regressió logística, aquest és un model basat en arbres de decisió i en una estratègia de boosting, on diversos arbres febles s'entrenen seqüencialment per corregir els errors dels anteriors. Aquesta naturalesa basada en arbres té implicacions directes en el preprocessament de les dades. Concretament, a diferència dels models basats en distàncies o en combinacions lineals de les variables (com el SVM o la regressió logística), no és necessari escalar les variables. Els arbres de decisió prenen decisions a partir de llindars sobre les variables, de manera que l'escala de les dades no afecta el procés de partició ni el rendiment del model. Per aquests motius, en el preprocessament associat a aquest model ens limitem a les transformacions estrictament necessàries (com ara la codificació de variables categòriques, si escau), però prescindim de l'escalat, ja que no aporta beneficis i no és requerit pel funcionament intern de l'algorisme.

## 5.1. Entrenament del model

En les primeres proves del model s'ha detectat un grau elevat d'overfitting, és a dir, el model s'ajustava massa bé a les dades d'entrenament però generalitzava malament sobre dades noves. Per aquest motiu, s'han aplicat diverses mesures amb l'objectiu de millorar la seva capacitat de generalització. En primer lloc, s'ha limitat la profunditat màxima dels arbres per evitar que siguin excessivament complexos i acabin memoritzant patrons molt específics del conjunt d'entrenament. A més, s'ha incrementat el paràmetre `min_child_weight`, que exigeix un nombre mínim de mostres per poder crear noves divisions, reduint així la probabilitat de generar splits basats en pocs exemples que no generalitzen bé. En segon lloc, s'han introduït mecanismes de regularització mitjançant els paràmetres `reg_alpha` (regularització L1) i `reg_lambda` (regularització L2), que penalitzen pesos excessivament grans i afavoreixen models més simples i robustos. També s'ha ajustat el paràmetre `gamma`, que defineix la reducció mínima de pèrdua necessària per fer una divisió, actuant com un cost addicional de complexitat.

Concretament, aquestes han sigut les combinacions d'hiperparàmetres provades per la Validació Creuada (5-fold CV), els hiperparàmetres seleccionats estan marcats en negreta:

| **Hiperparàmetre** | **Valors provats** |
|---|---|
| `n_estimators` | 50, 100, 150 |
| `max_depth` | 2, 3, 4 |
| `learning_rate` | 0.01, 0.05, **0.1** |
| `subsample` | **0.6**, 0.8 |
| `colsample_bytree` | 0.6, **0.8** |
| `min_child_weight` | **3**, 5, 7 |
| `gamma` | 0.1, 0.5, 1.0 |
| `reg_alpha` | 0.1, 0.5, 1.0 |

| Hiperparàmetre | Valors provats |
|---|---|
| `reg_lambda` | 1.0, 2.0, 5.0 |
| `scale_pos_weight` | 2, 3 |

Cal destacar el `learning_rate`, que controla la contribució de cada arbre nou al model final, actuant com un factor d'escala que pondera les prediccions individuals. S'han provat valors de 0.01, 0.05 i 0.1, un rang conservador que prioritza la convergència estable i la capacitat de generalització davant la velocitat d'entrenament. Després tenim els hiperparàmetres de mostreig. `subsample` controla la fracció d'observacions utilitzades per entrenar cada arbre, mentre que `colsample_bytree` defineix la proporció de característiques considerades. Els valors provats introdueixen suficient aleatorietat per reduir la correlació entre arbres i prevenir el sobreajust, sense limitar excessivament la informació disponible per captar patrons rellevants.

## 5.2. Avaluació del model

Un cop tenim tots els hiperparàmetres seleccionats, procedim amb l'avaluació del model. Comencem analitzant la matriu de confusió i la corba ROC del model sobre el conjunt de validació:

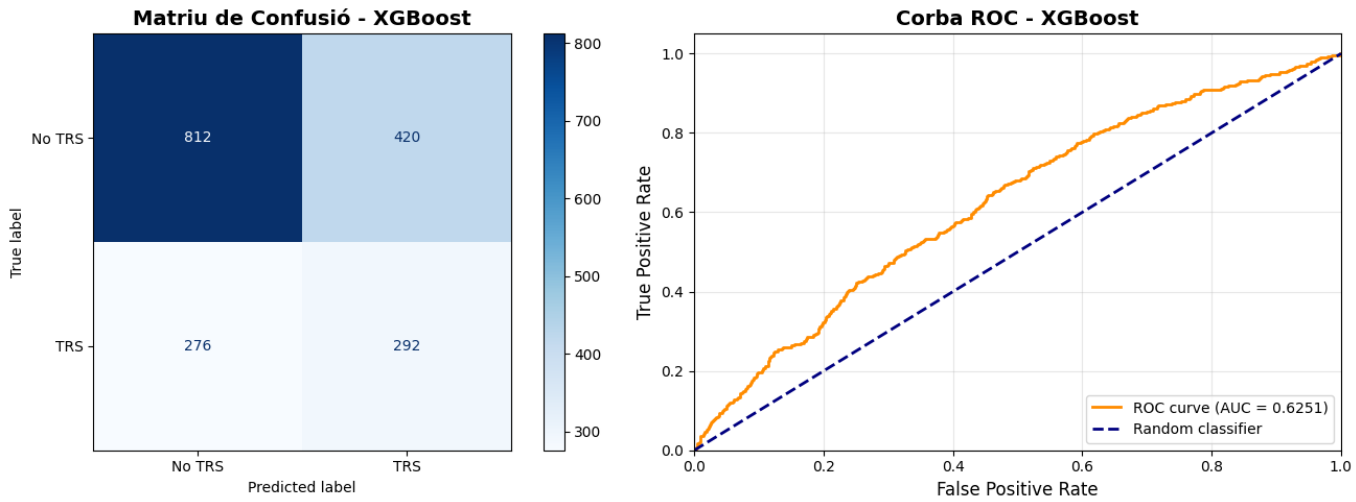



*Figura 9. Matriu de confusió i corba ROC del XGBoost (xgboost.ipynb).*

Observem que en aquest model es manté el desequilibri ja detectat en el model SVM. Pel que fa a la classe NO TRS, el model obté 812 veritables negatius i 420 falsos positius. En relació amb la classe TRS, es registren 276 falsos negatius i 292 veritables positius. Aquests resultats indiquen que el rendiment del model és limitat, ja que presenta un nombre elevat de falsos positius i una quantitat no negligible de falsos negatius. Les conclusions són coherents amb les exposades anteriorment: en funció de l'objectiu clínic perseguit, el comportament del model pot considerar-se més o menys adequat. En aquest sentit, seria recomanable valorar un canvi en la mètrica optimitzada durant el procés d'entrenament per ajustar millor el model a les necessitats clíniques. Pel que fa a la corba ROC, el model assoleix un valor d'AUC = 0,6251, un resultat relativament baix que indica una capacitat discriminativa limitada, propera a la que s'obtindria amb un model aleatori.

Analitzem ara les metriques del nostre model, tant per el conjunt d'entrenament com per el de validació, per veure si es capaç de generalitzar correctament i si pateix d'overfitting o underfitting.

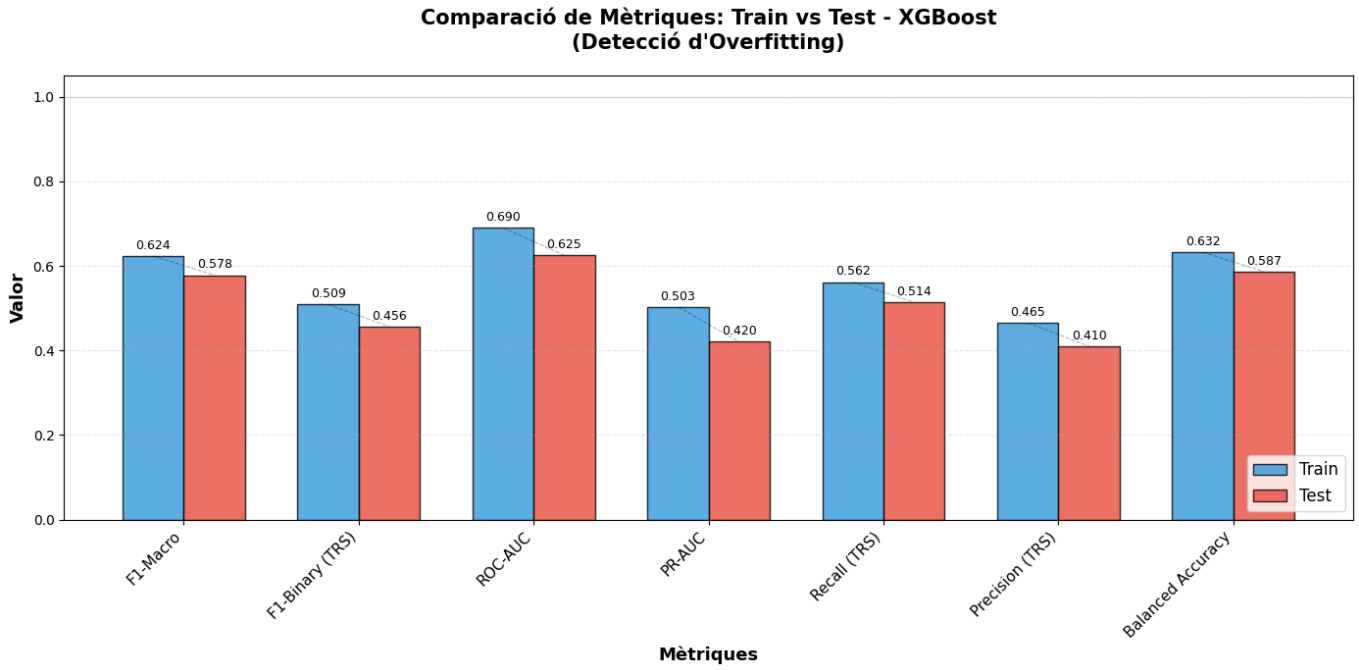



*Figura 10. Gràfics de barres de les mètriques del XGBoost (xgboost.ipynb).*

Les diferències entre els conjunts de train i test són consistentment baixes (majoria entre 0.04-0.06), demostrant que el model generalitza adequadament. Aquests gaps petits confirmen que les tècniques de regularització aplicades (max_depth reduït, min_child_weight elevat, reg_alpha i reg_lambda augmentats) han estat efectives per prevenir sobreajustament. El ROC-AUC de 0.625 al test indica una bona capacitat de separació entre classes. El PR-AUC de 0.420 és inferior, reflectint les dificultats inherents a datasets desbalancejats, on la precisió en la classe minoritària (TRS=1) és més difícil d'aconseguir. L'F1-Macro de 0.578 al test mostra un equilibri raonable entre precisió i recall per ambdues classes. L'F1-Binary específic per TRS (0.456) és més baix, consistent amb la dificultat de predir correctament la classe positiva. El model aconsegueix un recall de 0.514 per TRS al test, identificant aproximadament la meitat dels casos resistents. La precision de 0.410 indica que 4 de cada 10 prediccions positives són correctes, el que és un resultat bastant pobre.

En conclusió, el model XGBoost presenta un rendiment limitat però estable, amb una capacitat de generalització adequada gràcies a les tècniques de regularització aplicades, com ho demostren les diferències petites entre train i test (0.04-0.06 punts). Tot i això, els resultats absoluts revelen deficiències importants en la detecció de la classe TRS: només s'identifica correctament la meitat dels casos resistents (Recall = 0.514) i gairebé 6 de cada 10 prediccions positives són falsos positius (Precision = 0.410), un resultat poc satisfactori des d'un punt de vista clínic. El ROC-AUC de 0.625 i el PR-AUC de 0.420 confirmen aquesta capacitat discriminativa limitada, propera a un classificador aleatori, que reflecteix tant la complexitat inherent del problema com les dificultats associades al desbalanceig del dataset.

# 6. Regressió logística

Per a aquest model, ja no s'utilitzarà la implementació de sklearn, sino que es desenvolupa la classe de forma manual.

El model implementa una funció sigmoide per transformar les prediccions lineals en probabilitats. Per evitar overflow numèric, els valors d'entrada a la sigmoide es restringeixen a l'interval [-250, 250]. L'entrenament utilitza descens de gradient estocàstic per mini-batch amb les seguents característiques: els pesos s'actualitzen per lots de dades en lloc d'utilitzar tot el dataset, accelerant la convergència i permetent escapar de mínims locals; les dades es barregen a cada època per evitar que el model aprengui patrons basats en l'ordre de les dades; els pesos s'inicialitzen amb valors petits seguint una distribució normal. El model además incorpora ponderació de classes per compensar el desbalanceig del dataset. Els pesos es calculen inversament proporcionals a la frequencia de cada classe:

$$w_c = n_{samples} / (2 \cdot n_c)$$

Aquesta ponderació s'aplica a l'error durant el backpropagation, donant més importància als errors de la classe minoritària. Per últim, s'ha implementat regularització Ridge que penalitza valors grans dels pesos per evitar overfitting.

## 6.1. Entrenament del model

Per al desenvolupament de la regressió logística custom, aplicarem la mateixa estratègia de preprocessament que per el SVM. Al igual que en els models anteriors, farem una cerca d'hiperparàmetres per trobar la combinació que maximitza F1-macro. Aquesta és la graella on hem executat la cerca, i els valors en negreta són els escollits per validació creuada:

| Hiperparàmetre | Valors provats |
|---|---|
| `l2_reg` | **0.001**, 0.01, 0.1 |
| `learning_rate` | **0.001**, 0.005, 0.01, 0.1 |
| `batch_size` | 32, 64, 128, 256, **512** |
| `n_epochs` | 50, 100, **150**, 200 |

Per què s'han escollit aquests hiperparàmetres i valors? El `learning_rate` controla la magnitud dels passos que el model fa durant l'actualització dels pesos en cada iteració del descens del gradient, s'ha utilitzat una cerca logarítmica que cobreix tres ordres de magnitud per identificar el balanç òptim entre velocitat de convergència i estabilitat. El `batch_size` és el nombre de mostres que s'utilitzen en cada iteració per calcular el gradient abans d'actualitzar els pesos, s'han explorat potències de 2. Respecte `l2_reg`, és la penalització que s'afegeix a la funció de pèrdia proporcional al quadrat dels pesos, ajuda a prevenir l'overfitting. Per últim, `n_epochs` és el nombre de vegades que l'algorisme revisa tot el conjunt d'entrenament durant l'aprenentatge, i s'han explorat valors que cobreixen des d'entrenament ràpid fins a entrenament extens.

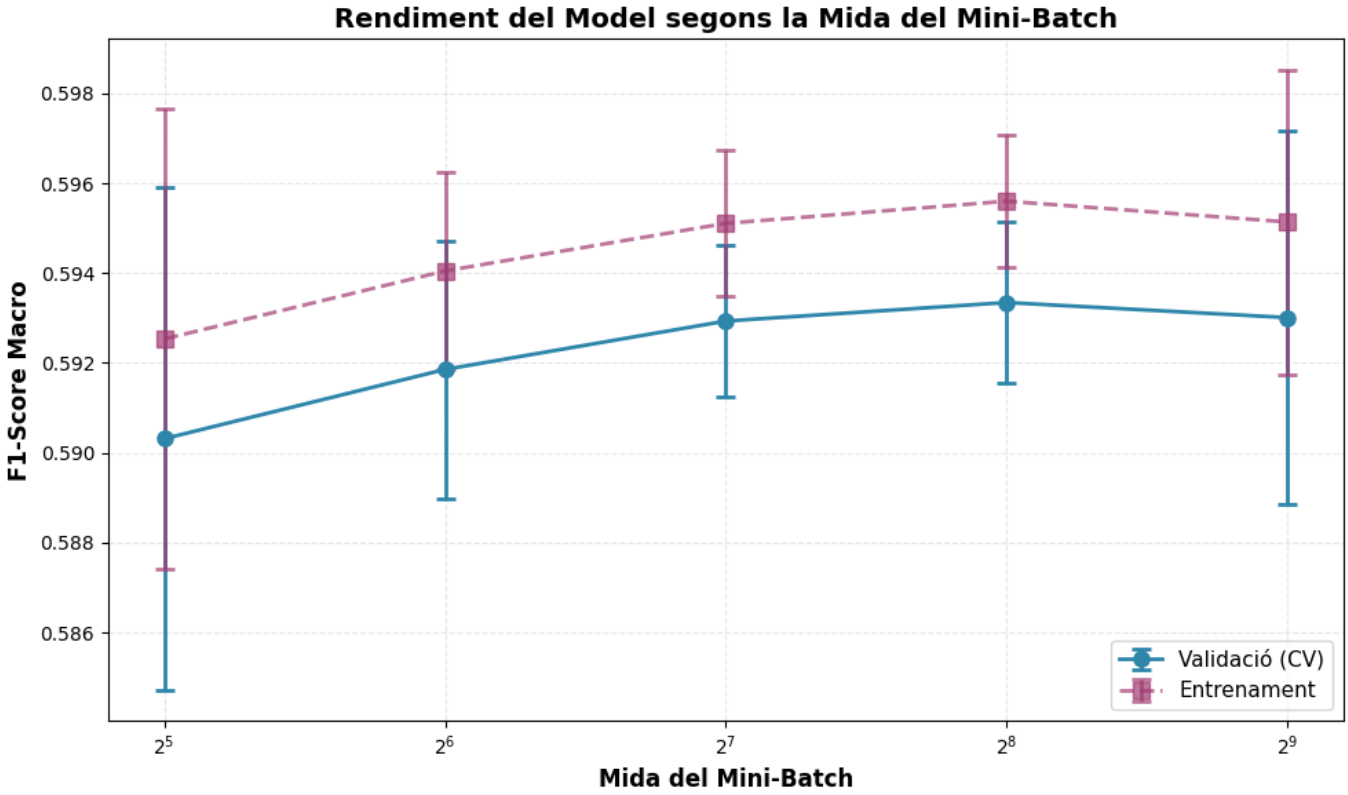



*Figura 11. F1-Macro en funció de la mida del mini-batch (logistic_regression.ipynb).*

En aquest gràfic s'observa el rendiment del model (F1-macro) segons la mida del batch, tant per el conjunt d'entrenament com per el de validació. Veiem que pels dos conjunts la tendència és manté igual, el rendiment creix de forma gradual fins a arribar a un pic en la mida de 256. Després, baixa una mica. Curiosament, la mida escollida en la validació creuada no ha sigut on el pic és màxim, sino en 512. Aquesta aparent contradicció s'explica perquè el gràfic mostra la mitjana agrupada de tots els resultats per cada batch_size, mentre que el GridSearchCV selecciona la millor combinació específica de tots els hiperparàmetres. És a dir, tot i que la mitjana de totes les configuracions amb `batch_size=256` és lleugerament superior, la combinació particular d'hiperparàmetres amb `batch_size=512` (juntament amb els seus valors específics de learning_rate, l2_reg i n_epochs) ha proporcionat el millor resultat individual.

## 6.2. Avaluació del model

Amb el model preparat, procedim amb la seva validació. Es comena observant la matriu de confusió i la corba ROC:

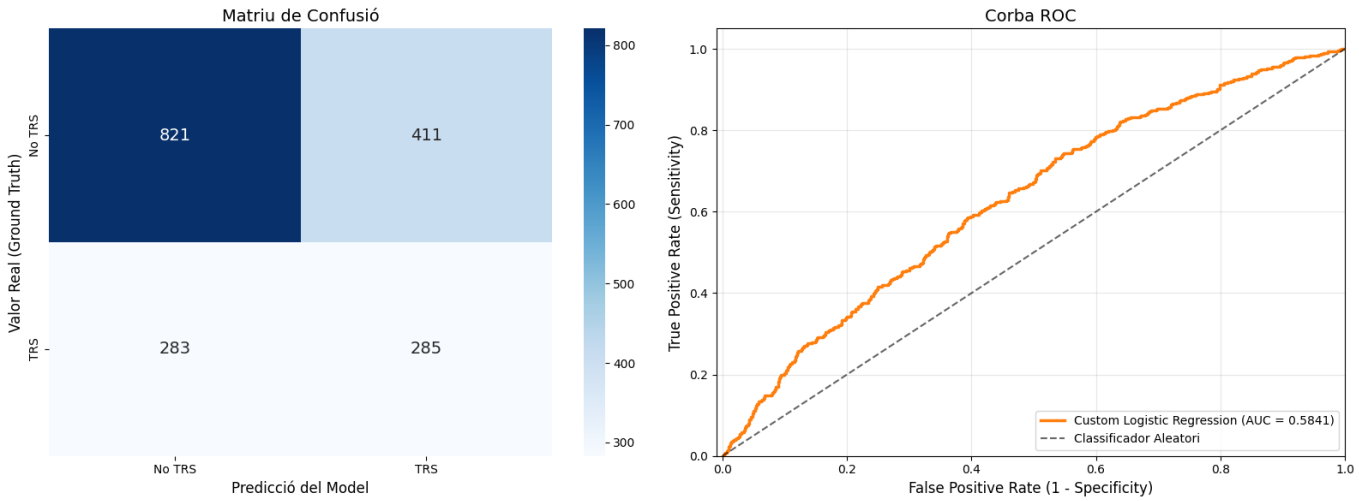



*Figura 12. Matriu de confusió i corba ROC de la regressió lineal (logistic_regression.ipynb).*

Observem que el patró observat en els dos models anteriors es manté. Per la classe NO TRS, tenim 821 veritables negatius i 411 falsos positius. Per a TRS, tenim 283 falsos negatius i 285 veritables positius. Respecte la corba ROC, veiem un AUC de 0.5841, similar al dels models anteriors.

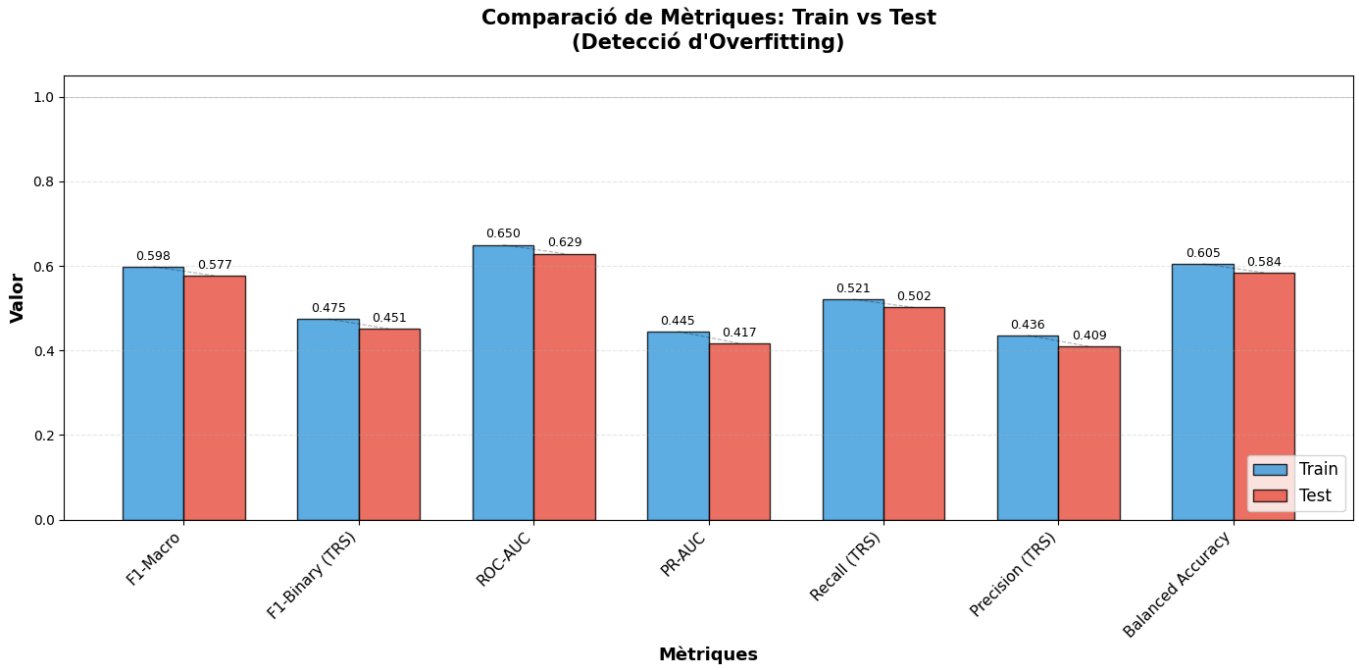



*Figura 13. Gràfics de barres de les mètriques de la regressió lineal (logistic_regression.ipynb).*

Comparant les mètriques d'avaluació en el conjunt de train i en test, podem analitzar la capacitat de generalització del model i detectar possibles problemes d'overfitting. Analitzant les diferències absolutes entre train i test, observem que són molt petites en totes les mètriques. Per exemple, en F1-Macro obtenim 0.598 en train i 0.577 en test, una diferència de només 0.021 punts. De manera similar, el ROC-AUC mostra valors de 0.650 en train i 0.629 en test, amb una diferència de 0.021. La mètrica F1-Binary (TRS) presenta valors de 0.475 en train i 0.451 en test (diferència de 0.024), mentre que el PR-AUC mostra 0.445 en train i

0.417 en test (diferència de 0.028). Les mètriques de Recall (TRS) i Precision (TRS) també mostren patrons similars. El Recall obté 0.521 en train i 0.502 en test (diferència de 0.019), mentre que la Precision presenta 0.436 en train i 0.409 en test (diferència de 0.027). Finalment, la Balanced Accuracy, que és especialment important en conjunts desbalancejats, mostra 0.605 en train i 0.584 en test, amb una diferència mínima de 0.021 punts. El fet que totes aquestes diferències siguin inferiors a 0.03 punts indica que el model no està memoritzant el conjunt d'entrenament, sinó que està aprenent patrons generalitzables. Si el model patís d'overfitting sever, esperaríem veure valors molt més alts en train (per exemple, F1-Macro > 0.9) amb caigudes substancials en test (per exemple, F1-Macro < 0.5). En canvi, el rendiment és consistent i comparable entre ambdós conjunts.

En conclusió, el model de Regressió Logística implementat manualment ofereix un rendiment comparable als altres models avaluats, amb un F1-Macro de 0.58 en validació que indica un equilibri raonable entre la detecció d'ambdues classes. Les diferències entre train i test es mantenen en un rang acceptable, confirmant que el model generalitza adequadament sense sobreajustament significatiu gràcies a la regularització L2 aplicada i al balanceig de pesos per classes. No obstant això, igual que en els models anteriors, la capacitat predictiva global és moderada, amb un ROC-AUC que indica dificultats per discriminar efectivament entre pacients amb TRS i sense TRS.

# 7. Selecció de model

Un cop tenim els tres models entrenats i validats, procedim amb la selecció del millor model. Per a escollir un, s'analitzen i es comparen les mètriques de cada un d'ells:

| Mètrica | SVM | XGBoost | Logistic Regression |
|---|---|---|---|
| Accuracy | **0.64** | 0.61 | 0.61 |
| F1-Score (Macro) | **0.59** | 0.58 | 0.58 |
| F1-Score (Weighted) | **0.64** | 0.62 | 0.62 |
| ROC-AUC | 0.6281 | 0.6251 | **0.6291** |
| Precision (No TRS) | 0.74 | **0.75** | 0.74 |
| Recall (No TRS) | **0.72** | 0.66 | 0.67 |
| F1-Score (No TRS) | **0.73** | 0.70 | 0.70 |
| Precision (TRS) | **0.43** | 0.41 | 0.41 |
| Recall (TRS) | 0.45 | **0.51** | 0.50 |
| F1-Score (TRS) | 0.44 | **0.46** | 0.45 |

El model SVM destaca clarament en les mètriques d'accuracy i F1-score weighted, obtenint un valor de 0.64 en ambdós casos, mentre que XGBoost i Logistic Regression assoleixen 0.61-0.62. Aquesta diferència de 0.033 punts en accuracy pot semblar petita, però és significativa en contextos clínics on cada predicció correcta pot influir en decisions terapèutiques. Pel que fa al F1-score macro, el SVM també lidera amb 0.59, superant lleugerament els altres dos models que obtenen 0.58. Respecte el ROC-AUC, la Regressió Logística obté el millor resultat amb 0.6291, seguida de molt a prop pel SVM (0.6281) i XGBoost (0.6251). Les diferències són mínimes (menys d'un punt percentual), indicant que tots tres models tenen una capacitat de discriminació similar. Ara, observem el rendiment per a cada classe:

- Per NO TRS, el SVM mostra el millor rendiment global per aquesta classe, amb un F1-score de 0.73. Aquest valor es deu principalment al seu excel·lent recall de 0.72, el més alt entre els tres models, tot i que XGBoost obté una precision lleugerament superior (0.75 vs 0.74). La capacitat del SVM per identificar correctament el 72% dels casos de No TRS és crucial per evitar falsos positius que podrien derivar en tractaments innecessaris.

- Per TRS, XGBoost obté els millors resultats, amb un F1-score de 0.46, impulsat pel seu recall de 0.51, significativament superior al del SVM (0.45) i Logistic Regression (0.50). Això significa que XGBoost és capaç de detectar el 51% dels casos reals de TRS, reduint els falsos negatius que podrien deixar pacients sense el tractament adequat. No obstant això, el SVM manté la millor precision per TRS (0.43), indicant menys falsos positius.

Després d'analitzar aquests resultats, es selecciona el SVM com el millor model. Aquest obté el millor rendiment de forma global, amb els valors més alts d'accuracy i F1-score weighted, indicant un rendiment superior quan es consideren ambdues classes amb els seus pesos. És veritat que XGBoost té un millor recall per TRS, però el SVM ofereix un millor equilibri global, amb un F1-score macro de 0.59, el més alt dels tres models i la mètrica que s'ha maximitzat durant l'entrenament. Ademés, és el que té un menor risc de falsos positius, amb una precicion de 0.43 per TRS.

Cal destacar que per a la selecció del model s'ha prioritzat el model amb un millor rendiment de forma global per les dues classes, és a dir, donem la mateixa importància tant a TRS com a NO TRS. Com ja s'ha mencionat anteriorment, els objectius clínics poden variar, i si volguéssim prioritzar una classe sobre una altra, l'entrenament i la selecció del model haurien sigut diferents. Per exemple, si el nostre objectiu és tenir el mínim nombre de falsos negatius, el XGBoost hagués sigut el model guanyador.

**Característiques del model triat**

El problema abordat consisteix a predir la presència o absència de esquizofrènia resistent al tractament (TRS) a partir d'un conjunt de variables clíniques, genètiques i de neuroimatge, amb una clara no linealitat potencial i un desbalanç entre les classes (No TRS vs TRS). En aquest context, el model SVM presenta un conjunt de característiques especialment adequades. Pel que fa a la complexitat, el SVM és capaç de capturar relacions no lineals entre les variables d'entrada i la probabilitat de TRS mitjançant la projecció de les dades a un espai de dimensió superior. Això permet modelar fronteres de decisió flexibles sense necessitat d'un feature engineering exhaustiu. Al mateix temps, els valors òptims trobats per als hiperparàmetres C i gamma situen al model en una zona de complexitat moderada, evitant tant el subajustament com el sobreajustament, tal com mostren les mètriques similars entre train i test. En termes d'interpretabilitat, el SVM no és tan transparent com la regressió logística, ja que no proporciona coeficients directament interpretables que relacionin cada variable amb el risc de TRS. Tot i això, la seva estructura segueix sent relativament compacta (basada en un conjunt de support vectors). Respecte als hiperparàmetres, el model utilitza principalment C i gamma, que controlen respectivament la penalització dels errors de classificació i el radi d'influència dels punts d'entrenament. Aquest nombre reduït d'hiperparàmetres facilita un procés d'ajust sistemàtic mitjançant validació creuada i permet entendre clarament els efectes de cada paràmetre sobre el comportament del model (més o menys flexibilitat de la frontera de decisió i més o menys tolerància a errors). Pel que fa al volum de dades, el nombre de mostres disponible és suficientment gran perquè un SVM pugui entrenar-se de manera fiable, però no tan elevat com per fer inviable el seu entrenament en temps raonable. El model és capaç de treballar eficientment amb el nombre actual d'observacions i la dimensionalitat del problema, tot mantenint uns temps de càlcul compatibles amb un escenari d'ús clínic o de recerca. En cas que el volum de dades creixés significativament en el futur, es podria estudiar l'ús de versions aproximades de SVM o d'altres models més escalables, però en l'estat actual del dataset el compromís entre capacitat de modelatge i cost computacional és favorable.

**Limitacions del model**

Tot i que el SVM presenta un rendiment global superior als altres models avaluats, és imprescindible reconèixer i analitzar les seves limitacions per entendre millor el seu abast i possibles millores futures. La limitació més crítica del model és el seu recall modest per a la classe minoritària (TRS), amb un valor de 0.45. Això significa que el model només identifica correctament el 45% dels casos reals de resistència al tractament, deixant un 55% de falsos negatius. En un context clínic, aquesta limitació és especialment preocupant, ja que els pacients amb TRS no detectats podrien seguir rebent tractaments convencionals ineficaços, retardant l'inici de teràpies més adequades i potencialment empitjorant el seu pronòstic. Aquest problema està parcialment relacionat amb el desbalanceig de classes del dataset. Com ja s'ha comentat, una altra limitació és la falta d'interpretabilitat. A diferència de la regressió logística, que proporciona coeficients interpretables per cada variable, el SVM no permet identificar fàcilment quines variables contribueixen més a la predicció de TRS. Ademés, el valor de ROC-AUC de 0.6281 indica que el model té una capacitat discriminativa acceptable però no excel·lent. Aquest valor suggereix que hi ha marge de millora significatiu, ja que un model perfecte obtindria un ROC-AUC d'1.0 i un classificador aleatori obtindria 0.5. La proximitat al valor aleatori indica que el model encara té dificultats per separar clarament les dues classes en molts casos, probablement degut a la complexitat inherent del problema i a possibles solapaments entre les distribucions de variables de pacients amb i sense TRS.

**Submission a Kaggle**

Ara que ja s'ha escollit el model que funciona millor, el fem servir per generar les prediccions sobre el conjunt `trs_eval` i poder enviar els resultats a la competició de Kaggle. Abans, però, apliquem exactament el mateix preprocessament que havíem fet servir amb `trs_train` (eliminem les mateixes variables, imputem els valors que falten, normalitzem les dades, etc.), per assegurar que el model vegi les dades amb el mateix format i pugui fer prediccions coherents.

Pel que fa al rendiment en el conjunt de competició, s'obté un F1-macro aproximat de 0.55, que es situa lleugerament per sota del valor assolit en el conjunt de validació intern, on el model havia arribat a un F1-macro al voltant de 0.585. Aquesta diferència és raonable i suggereix que el model generalitza de manera moderadament correcta: tot i que perd una mica de rendiment quan es troba amb dades completament noves, continua mostrant un comportament similar al que s'havia observat durant el procés de validació, sense caure en una degradació brusca que indicaria sobreajustament sever.

No obstant això, s'observa un comportament curiós quan es torna a entrenar el model optimitzant ROC-AUC en comptes de F1-macro. Tot i que aquest nou model obté un F1-macro inferior en el conjunt de test intern (com era d'esperar, ja que no està optimitzat per aquesta mètrica), quan es fa la predicció sobre el conjunt de Kaggle s'obté un F1-macro superior: 0.562. Això suggereix que el conjunt d'avaluació de Kaggle podria tenir una distribució lleugerament diferent a la del nostre conjunt de validació intern, o que l'optimització via ROC-AUC ha permès al model aprendre patrons que generalitzen millor sobre aquestes dades específiques. Aquest resultat reforça la importància d'explorar diferents mètriques d'optimització i ens recorda que el rendiment en validació creuada no sempre és un predictor perfecte del comportament en conjunts de dades externs.

# 8. Model Card for Treatment-resistant schizophrenia Diagnostic Dataset - SVM Classifier

## Model Details

### Overview

Aquest model prediu si una persona té esquizofrènia resistent al tractament o no, basant-se en les mesures clíniques donades. Aquest model està entrenat amb un algorisme de Support Vector Machines (SVM). És un algorisme que busca el hiperplà òptim que maximitza el marge de separació entre classes, utilitzant vectors de suport (punts més proxims a la frontera de decisió) per definir aquesta frontera. Els hiperparàmetres utilitzats són: paràmetre de regularització C (controla el trade-off entre maximitzar el marge i minimitzar els errors de classificació), gamma (defineix la influència de cada vector de suport en el kernel RBF) i el tipus de kernel (RBF per fronteresno lineals o lineal per separacions lineals)

### Version

name: 55d259d0-8010-4f45-b770-0f912fd0f235

date: 2025-12-27

### Owners

- Carlos Palazón Domingo(owner.role), carlos.palazon@estudiantat.upc.edu

### References

- https://www.kaggle.com/competitions/iaa-trs-detection
- https://scikit-learn.org/stable/modules/svm.html
- https://scikit-learn.org/stable/modules/preprocessing.html

## Considerations

### Intended Users

- Psiquiatres clínics amb experiència en el tractament de l'esquizofrènia que necessiten identificar precoçment pacients amb risc de resistència al tractament antipsicòtic. Aquests professionals utilitzen el model per complementar la seva avaluació clínica i optimitzar estratègies terapèutiques personalitzades.
- Equips multidisciplinaris de salut mental (psicòlegs clínics, infermeria especialitzada, treballadors socials) que participen en el seguiment i gestió de casos d'esquizofrènia. El model els ajuda a prioritzar recursos i intervencions per a pacients d'alt risc.
- Investigadors clínics i epidemiòlegs que estudien factors predictius de TRS i patrons de resposta al tractament, utilitzant el model per validar hipòtesis i generar noves línies de recerca en psiquiatria de precisió.

### Use Cases

- Aquest model està dissenyat com a eina de suport a la decisió clínica per identificar pacients amb esquizofrènia que presenten resistència al tractament antipsicòtic convencional (TRS). El model utilitza dades demogràfiques, clíniques, genètiques i de neuroimatge per proporcionar una estimació de risc de TRS. Els usuaris previstos són psiquiatres i professionals de salut mental en entorns hospitalaris o ambulatoris que necessitin orientació per ajustar estratègies terapèutiques. El model NO substitueix el criteri clínic professional i ha de ser utilitzat exclusivament com a complement al diagnòstic i seguiment mèdic. Les prediccions han de ser sempre validades per un especialista abans de prendre decisions sobre el tractament del pacient.

## Limitations

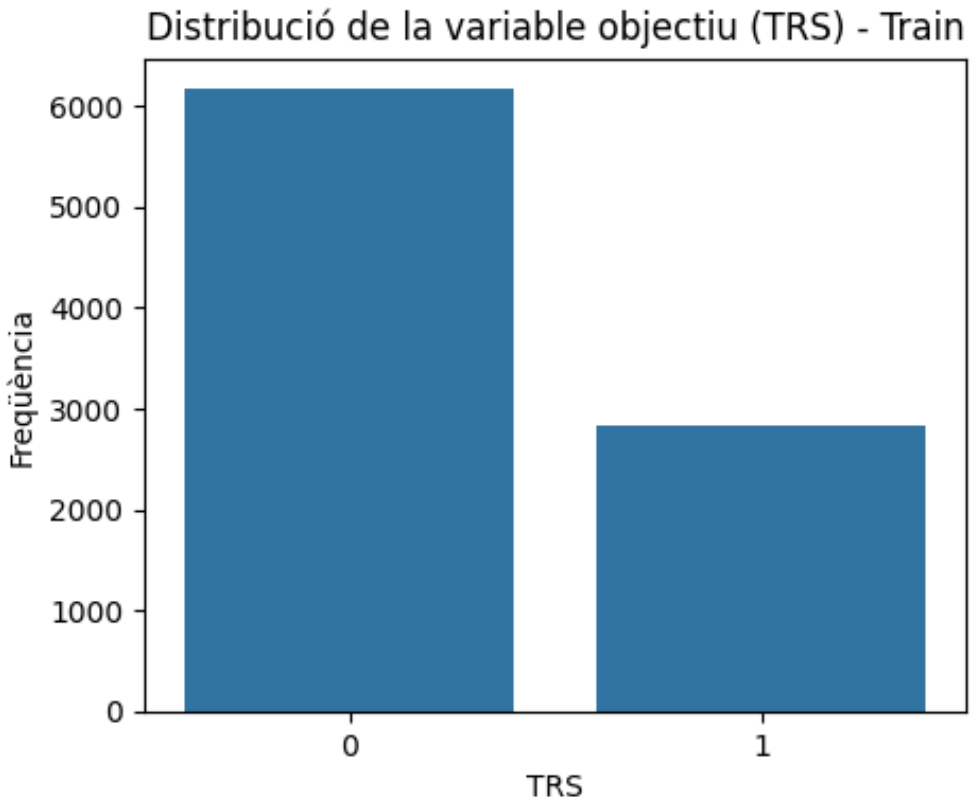
- Aquest model té limitacions tècniques conegudes:

1. Recall modest per TRS (0.45), identificant només el 0.45 dels casos reals de resistència, fet que pot retardar tractaments adequats en el 0.55 dels pacients no detectats.
2. ROC-AUC de 0.628 indica capacitat discriminativa limitada, propera a un classificador aleatori (0.5).
3. El model assumeix que les mostres d'entrenament són representatives de la població clínica general.
4. SVM amb kernel RBF manca d'interpretabilitat directa, dificultant identificar quines variables (genètiques, neuroimatge, clíniques) contribueixen més a la predicció.
5. El rendiment pot degradar-se amb dades de qualitat inferior o amb distribucions diferents de les observades durant l'entrenament.
6. La sensibilitat als hiperparàmetres C i gamma requereix reentrenament acurat si es modifica el dataset.

## Ethical Considerations

- **Risk:** Bias en les dades si determinats grups demogràfics estan subrepresentats
    - **Mitigation Strategy:** Analitzar rendiment del model per edat, gènere i origen ètnic, assegurant equitat entre subgrups
- **Risk:** El model pot perpetuar desigualtats si les dades reflecteixen accessibilitat desigual a tractaments
    - **Mitigation Strategy:** Validar que les etiquetes de resistència no estiguin confundides amb falta d'accés a tractament adequat
- **Risk:** Decisions clíniques basades només en prediccions automàtiques sense supervisió mèdica
    - **Mitigation Strategy:** El model és una eina de suport, no substitut del criteri clínic professional
- **Risk:** Privacitat i confidencialitat de dades mèdiques sensibles
    - **Mitigation Strategy:** Compliment amb anonimització de dades i protocols de seguretat robustos
- **Risk:** El model optimitzat amb F1-macro no prioritza específicament la minimització de falsos negatius
    - **Mitigation Strategy:** Avaluar recall de manera independent i considerar ajustar el threshold de decisió si volem el mínim nombre de falsos negatius

## Dataset




## Quantitative Analysis

- **Accuracy** - 0.595(train) - 0.587(test)
- **Precision** - 0.439(train) - 0.428(test)
- **Recall** - 0.462(train) - 0.452(test)
- **F1-Macro** - 0.594(train) - 0.585(test)
- **ROC-AUC** - 0.647(train) - 0.628(test)

Matriu de confusió i corba ROC en el test: