0. Taula de continguts

- 0. Taula de continguts
- 1. Introducció
- 2. Informació del problema
- 3. Anàlisi exploratori descriptiu
    - 3.1 Anàlisi de les variables
    - 3.2 Desbalanceig de la classe objectiu
    - 3.3 Valors perduts
    - 3.4 *Outliers*
    - 3.5 Estudi de dimensionalitat
    - 3.6 Partició del conjunt de dades
- 4. Ajustament de models
    - 4.1 SVM
        - 4.1.1 Preprocessament
        - 4.1.2 Ajustament del model
        - 4.1.3 Resultat final
    - 4.2 XGBoost
        - 4.2.1 Preprocessament
        - 4.2.2 Ajustament del model
        - 4.2.3 Resultat final
    - 4.3 Regressió logística personalitzada
        - 4.3.1 Preprocessament
        - 4.3.2 Ajustament del model
        - 4.3.3 Resultat final
- 5. Model final
- 6. Model Card
- 7. Conclusions

1. Introducció

2. Informació del problema

3. Anàlisi exploratori descriptiu

En aquesta secció es presenta l'anàlisi exploratori descriptiu (AED) realitzat sobre el conjunt de dades proporcionat per al problema de predicció clínica. L'objectiu principal d'aquesta anàlisi és comprendre millor les característiques de les dades, identificar possibles problemes com valors perduts o *outliers*, i preparar les dades per a l'ajustament dels models predictius.

3.1 Anàlisi de les variables

Observem primer de tot la taula de dades per tenir una visió general de les variables disponibles i la seva tipologia. L'obtenim amb la comanda `df.describe().T`, que ens proporciona estadístiques descriptives per a les variables numèriques i categòriques. El resultat ha estat:

Variable	N	Mean (SD)	Min	25%	Median	75%	Max	Missing (%)
Age	9000	26.04 (10.01)	13.00	19.00	25.00	31.00	64.00	0.00%
Sex (0=F, 1=M)	9000	0.58 (0.49)	0.00	0.00	1.00	1.00	1.00	0.00%
BMI	9000	28.11 (5.43)	15.00	24.40	28.00	31.70	49.60	0.00%
Duration Untreated Psychosis	8872	19.22 (19.55)	0.30	6.40	12.50	24.30	125.00	1.42%
Family History	9000	0.12 (0.32)	0.00	0.00	0.00	0.00	1.00	0.00%
Initial Response	9000	41.84 (30.16)	0.00	10.10	38.20	72.30	100.00	0.00%
Prior Antipsychotics	9000	0.41 (0.67)	0.00	0.00	0.00	1.00	2.00	0.00%
TRS (Target)	9000	0.32 (0.46)	0.00	0.00	0.00	1.00	1.00	0.00%
Lymphocyte count	7009	1.80 (0.60)	0.50	1.38	1.80	2.20	4.02	22.12%
Neutrophil count	7015	5.01 (1.47)	1.50	4.01	5.02	6.01	9.96	22.06%
Triglycerides	6547	152.01 (61.10)	40.00	108.05	151.10	194.60	394.60	27.26%
Glucose	6381	95.86 (18.31)	65.00	82.20	95.50	108.30	159.60	29.10%
Alkaline Phosphatase	6062	85.17 (24.83)	30.00	68.20	84.70	101.90	179.30	32.64%
IL-17A	8999	2.66 (0.80)	-0.20	2.12	2.65	3.21	5.38	0.01%
CCL23	9000	3.78 (1.05)	-0.20	3.08	3.78	4.49	7.69	0.00%
TWEAK	9000	4.19 (1.24)	-0.54	3.37	4.19	5.04	8.92	0.00%
HLA-DRB1*04:02	9000	0.02 (0.15)	0.00	0.00	0.00	0.00	1.00	0.00%
HLA-B*15:02	9000	0.03 (0.18)	0.00	0.00	0.00	0.00	1.00	0.00%
HLA-A*31:01	9000	0.05 (0.21)	0.00	0.00	0.00	0.00	1.00	0.00%
Polygenic Risk Score	8999	0.030 (0.14)	-0.44	-0.07	0.02	0.11	0.58	0.01%
Del 22q11.2	9000	0.009 (0.09)	0.00	0.00	0.00	0.00	1.00	0.00%
Ki Whole Striatum	9000	0.0130 (0.002)	0.0080	0.0113	0.0129	0.0145	0.0200	0.00%
Ki Associative Striatum	9000	0.0130 (0.002)	0.0071	0.0113	0.0128	0.0146	0.0210	0.00%
SUVRc Whole Striatum	9000	1.18 (0.27)	0.80	0.97	1.16	1.36	2.00	0.00%
SUVRc Assoc. Striatum	9000	1.18 (0.27)	0.80	0.96	1.16	1.37	2.00	0.00%

Taula 1: Resum estadístic de totes les variables numèriques i categòriques del conjunt d'entrenament. Es mostra el recompte (N), mitjana i desviació estàndard, quartils i percentatge de valors perduts.

Obeservem que el conjunt de dades conté un total de 9000 mostres i diverses variables predictives, així com la variable objectiu `TRS (Target)`. Algunes variables presenten valors perduts, especialment les mèdiques com el recompte de limfòcits i neutròfils, triglicèrids, glucosa i fosfatasa alcalina, que gestionarem més endavant a la secció 3.3.

Pel que fa les distribucions de les variables, la majoria semblen tenir una distribució aproximadament normal, o de combinacions de normals, com per exemple:

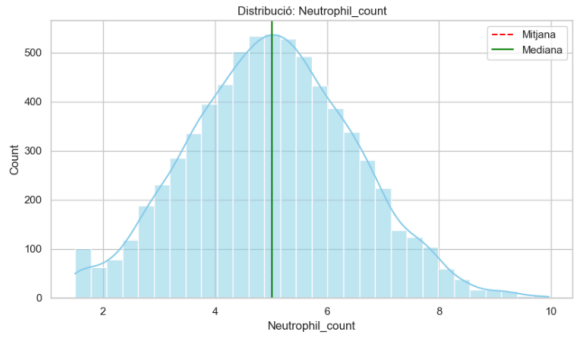

Figura 1: Distribució de `Neutrophil count`.

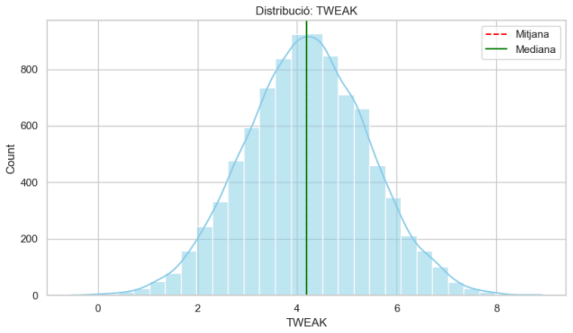

Figura 2: Distribució de `TWEAK`.

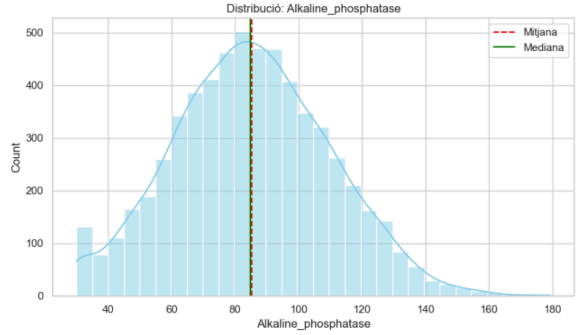

Figura 3: Distribució de `Alkaline phosphatase`

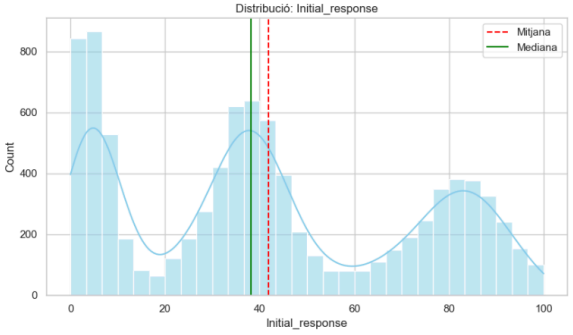

Figura 4: Distribució de `Initial response`.

Aquestes variables es poden utilitzar directament en els models predictius, ja que no requereixen transformacions addicionals per a la seva normalització. No obstant això, algunes variables com el `Duration_Untreated_Psychosis` o `Age` mostren una distribució més asimètrica:

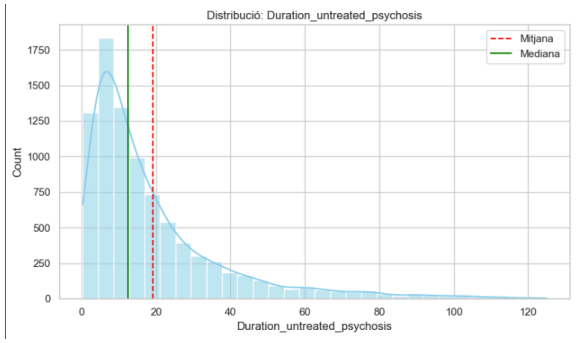

Figura 5: Distribució de `Duration_Untreated_Psychosis`.

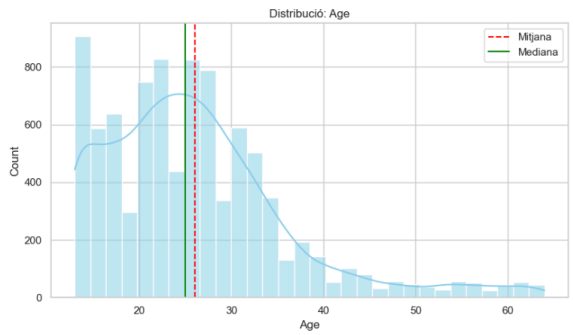

Figura 6: Distribució de `Age`.

En aquest cas podem veure unes cues llargues cap a la dreta, indicant la presència d'alguns valors extrems. Comprovem ràpidament amb el skewness que aquestes variables no són normals. Si tenen un skewness major a 1 o menor a -1, es consideren altament asimètriques. Comprovem:

```
ANÀLISI D'ASIMETRIA (SKEWNESS):
--------------------------------------------------------------------------------
                  Variable  Skewness  Kurtosis                                Estat
Duration_untreated_psychosis  2.150479  5.323006 🔴 MOLT ASIMÈTRICA (Requereix Log/BoxCox)
                       Age  1.258793  2.065980 🔴 MOLT ASIMÈTRICA (Requereix Log/BoxCox)
  SUVRc_associative_striatum  0.457536 -0.422680                    🟢 NORMAL (Simètrica)
        SUVRc_whole_striatum  0.434121 -0.408815                    🟢 NORMAL (Simètrica)
        Polygenic_risk_score  0.405186  0.323084                    🟢 NORMAL (Simètrica)
     Ki_associative_striatum  0.328552 -0.022675                    🟢 NORMAL (Simètrica)
...
```

Aquest desequilibri en la distribució de les dades pot afectar el rendiment dels models predictius, especialment aquells que assumeixen normalitat en les variables. Per tant, considerarem aplicar

transformacions com el logaritme o Box-Cox a aquestes variables abans de l'ajustament dels models.

Mirem la correlació entre les variables numèriques per identificar possibles relacions lineals que puguin ser útils per a la predicció. Utilitzem un mapa de calor per visualitzar aquestes correlacions:

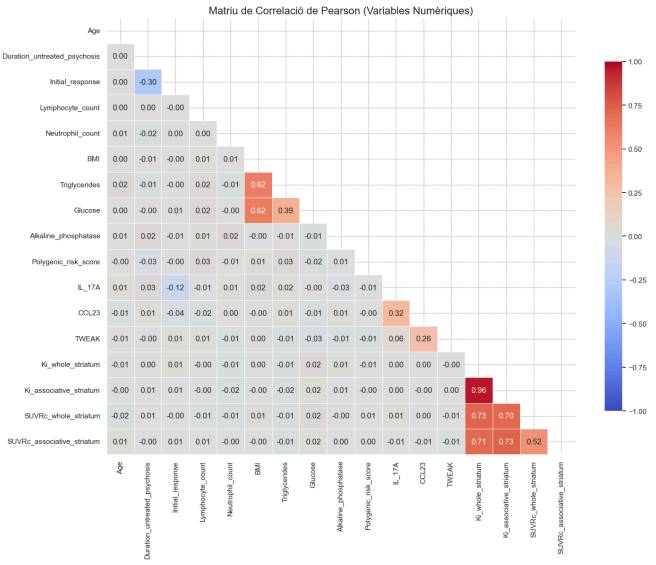

Figura 7. Matriu de Correlació de Pearson

Veiem que en termes generals, les correlacions entre les variables molt baixes o pràcticament 0, amb algunes excepcions:

Aquesta imatge mostra una **matriu de correlació de Pearson** (heatmap), que mesura la relació lineal entre diferents variables numèriques del teu dataset. Els valors van de **-1** (blau fosc, correlació negativa perfecta) a **+1** (vermell fosc, correlació positiva perfecta), passant per **0** (gris/blanc, sense correlació).

Aquí tens la interpretació detallada dels patrons més rellevants:

1. Redundància a les Variables avall dreta (Vermell Fosc) Aquest és el grup més destacat de la gràfica. Les variables relacionades amb l'estriat (*striatum*) mostren correlacions extremadament altes.
    - **`Ki_whole_striatum`** vs. **`Ki_associative_striatum`:** 0.96
    - **`Ki`** vs. **`SUVRc`**: ~0.70 - 0.73
2. Clúster Metabòlic (Vermell Mig):
    - **`Triglycerides`** vs. **`Glucose`**: 0.62
    - **`Glucose`** vs. **`BMI`**: 0.39
3. Relació entre no tratament i resposta inicial (Blau Fosc):
    - **`Duration_untreated_psychosis`** vs. **`Initial_response`**: -0.30

Aquestes correlacions suggereixen que certes variables estan fortament relacionades i podrien ser redundants en els models predictius. Gestionarem aquestes relacions en la fase de selecció de característiques per optimitzar el rendiment dels models.

3.2 Desbalanceig de la classe objectiu

La variable objectiu és `TRS`, que indica si un pacient desenvolupa resistència als tractaments antipsicòtics. La proporció de pacients la mirem executant `ratio_trs = df['TRS'].value_counts(normalize=True)`. El resultat és que el percentatge d'individus a la dataset que desenvolupa TRS és aproximadament del 31.5%, mentre que el 68.4% restant no desenvolupa resistència. Això indica un desbalanceig en les classes, que haurem de gestionar a l'hora de modelar. Aquest desbalanceig es denota en les variables categòriques:

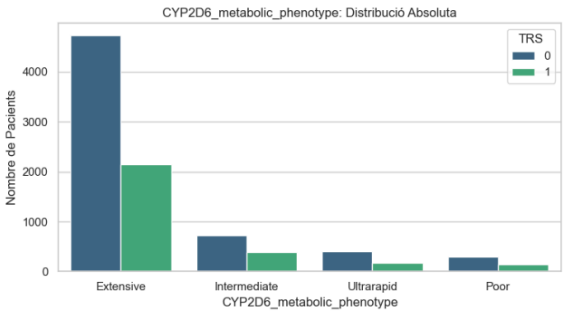

Figura 8: Distribució de la variable objectiu TRS

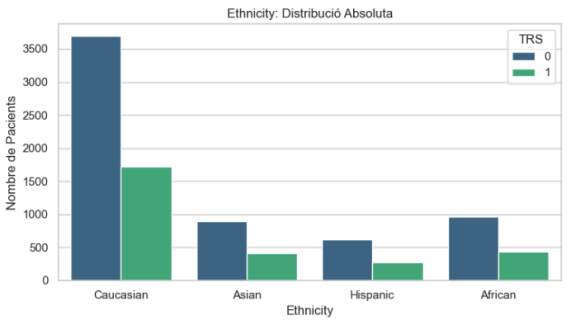

Figura 9: Distribució de la variable objectiu TRS

Veiem no hi ha cap biaix que provoqui el desbalanceig, ja que la distribució de les altres variables categòriques és força equilibrada. Per tant, per gestionar aquest desbalanceig en la fase d'ajustament dels models, considerarem tècniques com assignar pesos a les classes.

3.3 Valors perduts

Com podem veure a la Taula 1, algunes variables tenen un percentatge significatiu de valors perduts. Veiem-ho en més detall:

Variable	Total Missings	Percentatge (%)
Alkaline_phosphatase	2938	32.64%
Glucose	2619	29.10%
Triglycerides	2453	27.26%
Lymphocyte_count	1991	22.12%
Neutrophil_count	1985	22.06%
Duration_untreated_psychosis	128	1.42%
Polygenic_risk_score	1	0.01%
IL_17A	1	0.01%

Taula 2: Valors perduts per variable.

Experimentalment he provat d'imputar els valors perduts amb diferents tècniques, com la mitjana, mediana, KNN i regressió. Després d'avaluar el rendiment dels models amb aquestes diferents imputacions, he observat que la imputació mitjançant KNN amb k=5 ofereix els millors resultats en termes de precisió i robustesa del model. Per tant, he decidit utilitzar aquesta tècnica per gestionar els valors perduts en les variables mèdiques.

3.4 *Outliers*

He fet servir el mètode del IQR per detectar valors extrems en les variables numèriques. Aquest mètode defineix els *outliers* com aquells valors que es troben fora de l'interval $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$, on $Q1$ i $Q3$ són el primer i tercer quartil, respectivament, i IQR és el rang interquartílic (Q3 - Q1). La taula resultant mostra el nombre d'*outliers* detectats per variable:

Variable	n outliers	Outliers (%)
Duration_untreated_psychosis	651	7.23
Age	360	4.00
Polygenic_risk_score	125	1.39
Ki_associative_striatum	83	0.92
Ki_whole_striatum	71	0.79
CCL23	59	0.66
TWEAK	58	0.64
IL_17A	57	0.63
BMI	39	0.43
SUVRc_whole_striatum	35	0.39
Neutrophil_count	32	0.36
SUVRc_associative_striatum	29	0.32
Lymphocyte_count	26	0.29
Alkaline_phosphatase	23	0.26
Triglycerides	22	0.24
Glucose	16	0.18

Taula 3: Variables amb valors extrems (outliers).

Com veiem, la variable amb més outliers és `Duration_untreated_psychosis`, amb un total de 651 valors extrems, representant el 7.23% del total de mostres. Aquesta variable mostra una distribució molt asimètrica, amb una cua llarga cap a la dreta, indicant que hi ha alguns pacients amb períodes molt llargs sense tractament. Aquesta variable podria beneficiar-se d'una transformació logarítmica per reduir l'impacte dels outliers en els models predictius. El mateix passa amb la variable `Age`, que també presenta una quantitat significativa d'outliers (360 valors, 4.00%). La resta de variables tenen un nombre relativament baix d'outliers, tots per sota de l'1% del total de mostres.

Per tant, he decidit no eliminar aquests outliers, ja que podrien contenir informació rellevant sobre pacients amb característiques extremes. En lloc d'això, aplicaré transformacions adequades a aquestes variables per minimitzar el seu impacte en els models predictius. Pel que fa a les altres variables amb pocs outliers, no aplicaré cap acció específica, ja que la seva presència és mínima i no afectarà significativament els resultats dels models. A més, en un context mèdic, eliminar valors extrems podria conduir a la pèrdua d'informació important sobre pacients amb condicions rares o greus, que poden ser crucials per a la predicció.

3.5 Estudi de dimensionalitat

He realitzat una anàlisi de components principals (PCA) per avaluar la dimensionalitat del conjunt de dades i identificar possibles reduccions de dimensions que puguin millorar l'eficiència dels models predictius. La PCA és una tècnica estadística que transforma les variables originals en un nou conjunt de variables no correlacionades, anomenades components principals, que capturen la major part de la variància present en les dades. Com que no tolera valors perduts, he utilitzat el conjunt de dades amb els valors imputats mitjançant KNN. També eliminem la variable objectiu `TRS` abans d'aplicar la PCA, ja que aquesta tècnica només s'aplica a les variables predictives i la variable `id_patient`, que és un identificador únic per a cada pacient i no aporta informació rellevant per a la predicció. L'apliquem fent ús de la lliberieria `sklearn.decomposition.PCA`, i els resultats obtinguts són els següents:

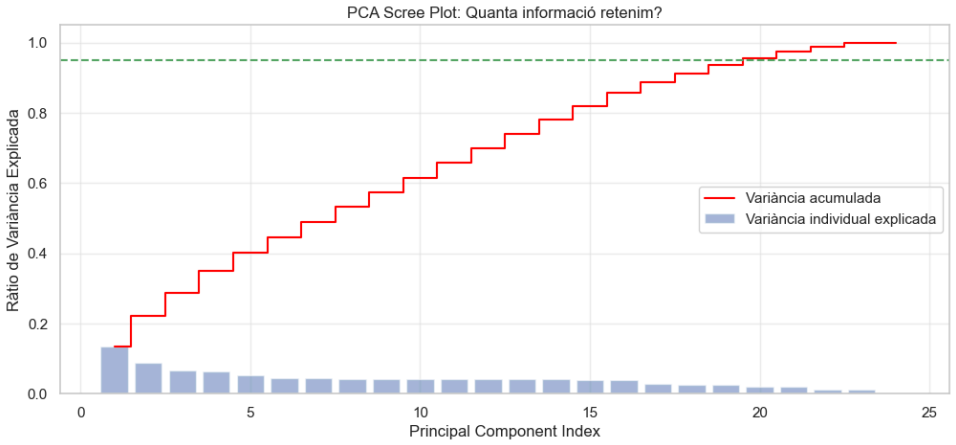

Figura 10: Gràfica de Scree Plot de la PCA

Podem veure que calen 20 components per explicar el 95% de la variància total del conjunt de dades. Això indica que hi ha una certa redundància entre les variables originals, ja que només es necessiten 20 components per capturar la major part de la informació. No obstant això, com que el nombre original de variables no és molt alt, he decidit no reduir la dimensionalitat en aquesta fase i mantenir totes les variables originals per a l'ajustament dels models. Això permetrà als models aprofitar tota la informació disponible i potencialment millorar el rendiment predictiu.

3.6 Partició del conjunt de dades

4. Ajustament de models

4.1 SVM

4.1.1 Preprocessament

4.1.2 Ajustament del model

4.1.3 Resultat final

4.2 XGBoost

4.2.1 Preprocessament

4.2.2 Ajustament del model

4.2.3 Resultat final

4.3 Regressió logística personalitzada

4.3.1 Preprocessament

4.3.2 Ajustament del model

4.3.3 Resultat final

5. Model final

6. Model Card

7. Conclusions