Informe de la pràctica

Cai Selvas Sala

13 de desembre de 2023

Universitat Politècnica de Catalunya
Grau en Intel·ligència Artificial
Introducció a l'Aprenentatge Automàtic

Resum

Aquest és l'informe corresponent a la pràctica individual de l'assignatura d'Introducció a l'Aprenentatge Automàtic del Grau en Intel·ligència Artificial de la Universitat Politècnica de Catalunya (UPC).

En el document s'explicarà com s'ha realitzat el projecte, quines dificultats s'hi han trobat, quins anàlisis s'han realitzat, quins resultats s'han obtingut i quines conclusions se'n poden extreure. A més, s'explica el codi en Python utilitzat durant la pràctica, així com els detalls del model creat (model card).

Índex

1 Introducció

1.1 Base de dades

[1]

1.2 Descripció del projecte

2 Documents i estructura general de l'entrega

3 Anàlisis i preprocessat de dades

3.1 Preprocessat inicial

Una vegada importem el dataset, es pot veure que hi ha bastantes cel·les buides i altres amb el string `'NaNN'`. Per solucionar aquesta inconsistència, s'han reemplaçat tots aquests valors per `pd.NA`.

Per altra banda, s'ha declarat el tipus de cada variable correctament (com a numèriques o com a categòriques) seguint la informació que es proporciona en el metadata file (que es pot trobar en [1]).

Addicionalment, per una millor comprensió de la variable *Status*, s'ha decidit reanomenar els seus valors, tenint en compte el metadata file, de la següent manera:

- **'C'** → 'Alive'.
- **'CL'** → 'Liver Transplant'.
- **'D'** → 'Dead'.

Una vegada realitzats aquests canvis, es pot començar a treballar amb el dataset correctament.

3.2 Anàlisis estadístic de les variables i estudi de balanceig de classes

El primer que s'ha fet per entendre el dataset i poder treballar amb ell és realitzar un anàlisis estadístic de cada una de les variables que el formen. A més, per les variables numèriques podem analitzar la distribució que segueixen mitjançant un histograma, mentre que per les categòriques podem realitzar countplots per veure la distribució entre les seves classes i com de balancejades estan.

3.2.1 Variables numèriques

En les taules 1 i 2 es poden veure estadístiques sobre totes les variables numèriques del dataset (obtingudes mitjançant la comanda `data.describe()` de la llibreria `pandas`).

	N_Days	**Age**	**Bilirubin**	**Cholesterol**	**Albumin**
count	418.0	418.0	418.000000	284.0	418.000000
mean	1917.782297	18533.351675	3.220813	369.510563	3.497440
std	1104.672992	3815.845055	4.407506	231.944545	0.424972
min	41.0	9598.0	0.300000	120.0	1.960000
25%	1092.75	15644.5	0.800000	249.5	3.242500
50%	1730.0	18628.0	1.400000	309.5	3.530000
75%	2613.5	21272.5	3.400000	400.0	3.770000
max	4795.0	28650.0	28.000000	1775.0	4.640000

Taula 1: Estadístiques de les variables numèriques N_Days, Age, Bilirubin, Cholesterol i Albumin.

	Copper	Alk_Phos	SGOT	Triglycerides	Platelets	Prothrombin
count	310.0	312.000000	312.000000	282.0	407.0	416.000000
mean	97.648387	1982.655769	122.55364	124.702128	257.02457	10.731731
std	85.61392	2140.388824	56.699525	65.148639	98.325585	1.02200
min	4.0	289.000000	26.350000	33.0	62.0	9.00000
25%	41.25	871.500000	80.600000	84.25	188.5	10.00000
50%	73.0	1259.000000	114.700000	108.0	251.0	10.60000
75%	123.0	1980.000000	151.900000	151.0	318.0	11.10000
max	588.0	13862.400000	457.250000	598.0	721.0	18.00000

Taula 2: Estadístiques sobre les variables numèriques Copper, Alk_Phos, SGOT, Triglycerides, Platelets i Prothrombin

Addicionalment, en les figures 1 i 2 es poden veure les histogrames de cada una de les variables numèriques, on es veu la distribució de les seves dades ignorant els valors faltants (missings).

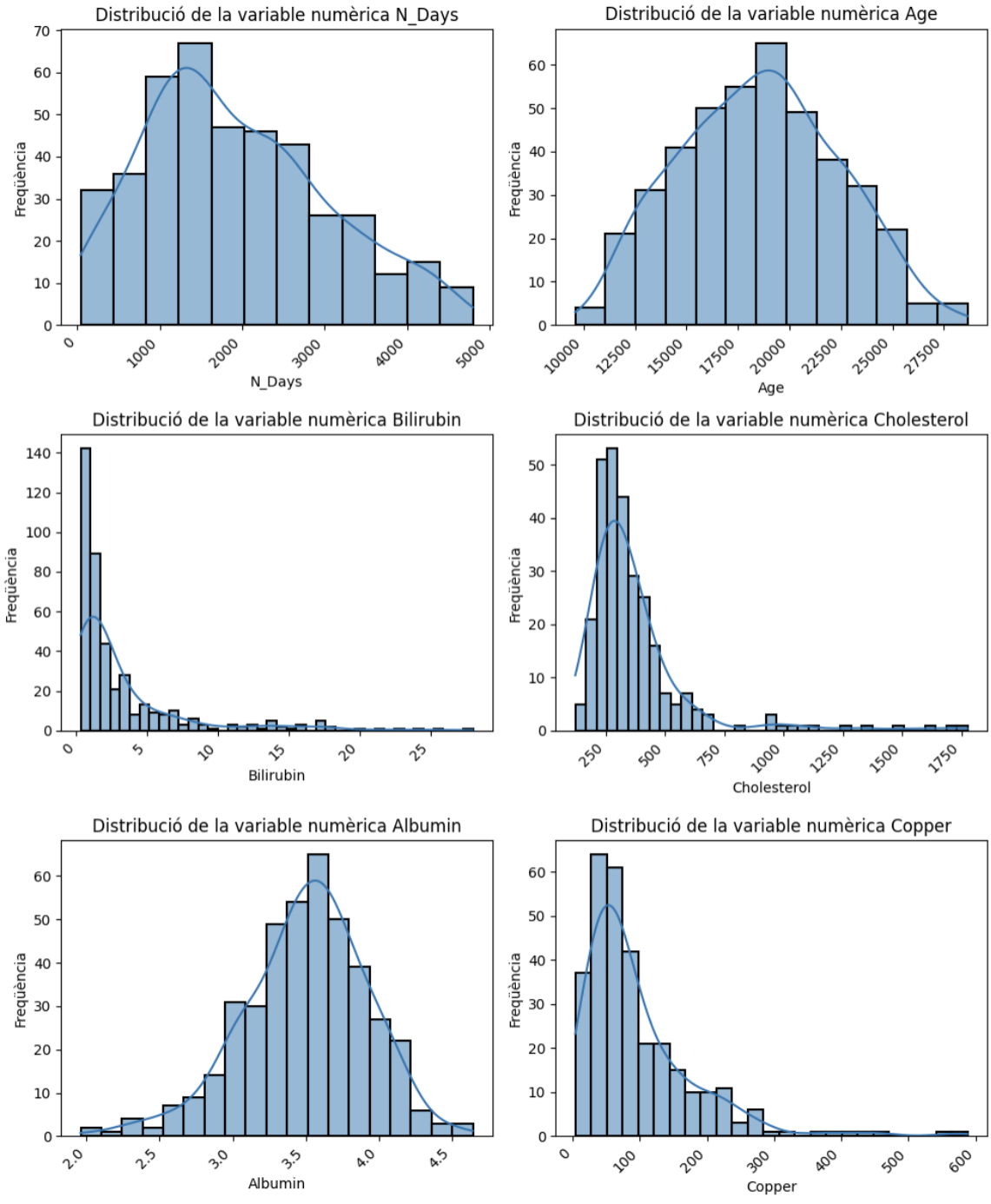

Figura 1: Histogrames de variables numèriques del datset.

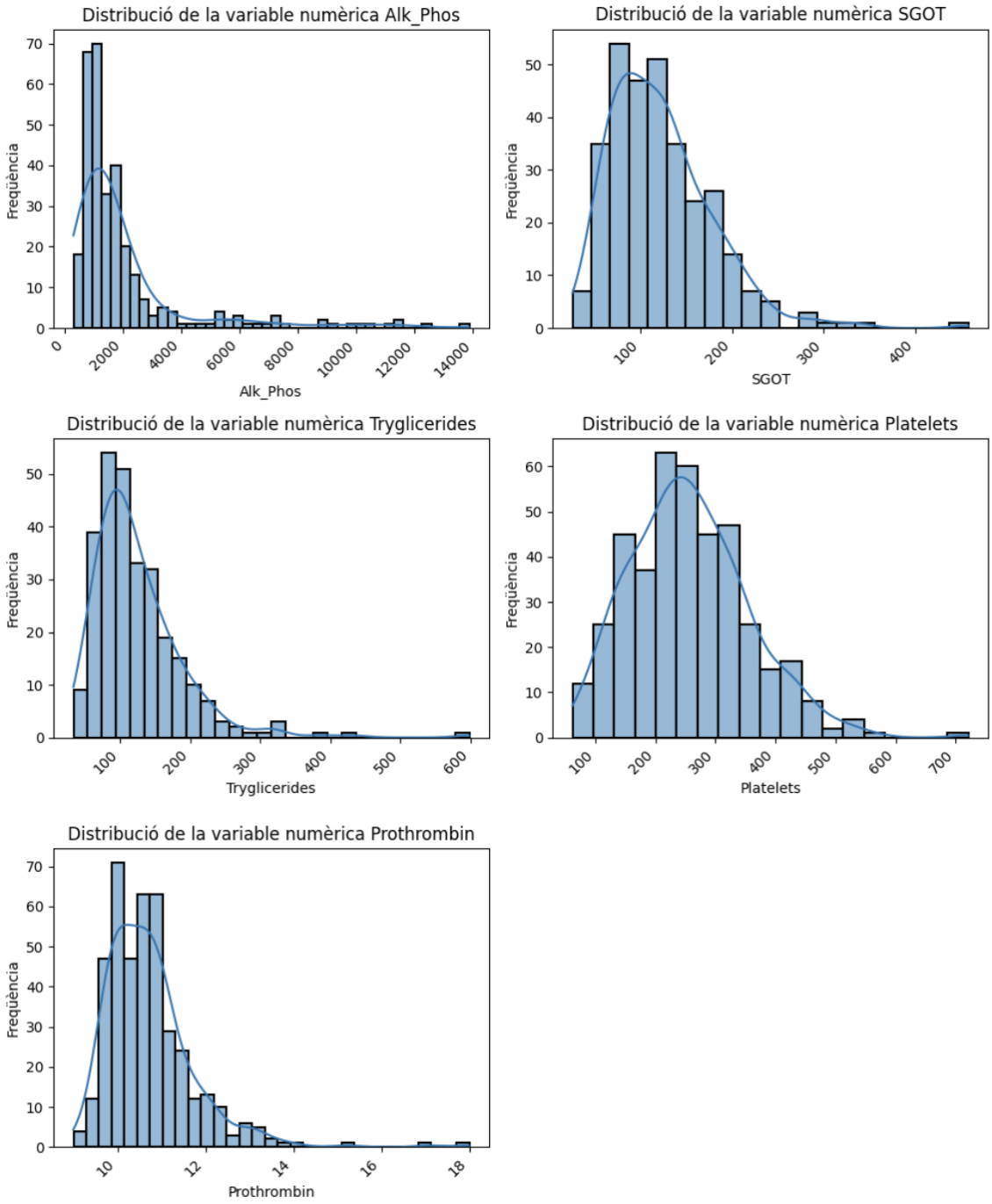

Figura 2: Histogrames de variables numèriques del datset.

3.2.2 Variables categòriques

En les taules 3 i 4 podem veure altres estadístiques per les variables categòriques del dataset (obtingudes mitjançant la mateixa comanda, però amb el paràmetre `include='category'`).

	ID	**Status**	**Drug**	**Sex**	**Ascites**
count	418	418	312	418	312
unique	418	3	2	2	2
top	1	Alive	D-penicillamine	F	N
freq	1	232	158	374	288

Taula 3: Estadístiques sobre les variables categòriques ID, Status, Drug, Sex i Ascites.

	Hepatomegaly	**Spiders**	**Edema**	**Stage**
count	312	312	418	412.0
unique	2	2	3	4.0
top	Y	N	N	3.0
freq	160	222	354	155.0

Taula 4: Estadístiques sobre les variabes categòriques Hepatomegaly, Spiders, Edema i Stage.

A més, en les figures 3 i 4 es poden veure els countplots de cada una de les variables categòriques, on es veu la quantitat de mostres que hi ha per cada classe de la variable, evitant els valors faltants (missings).

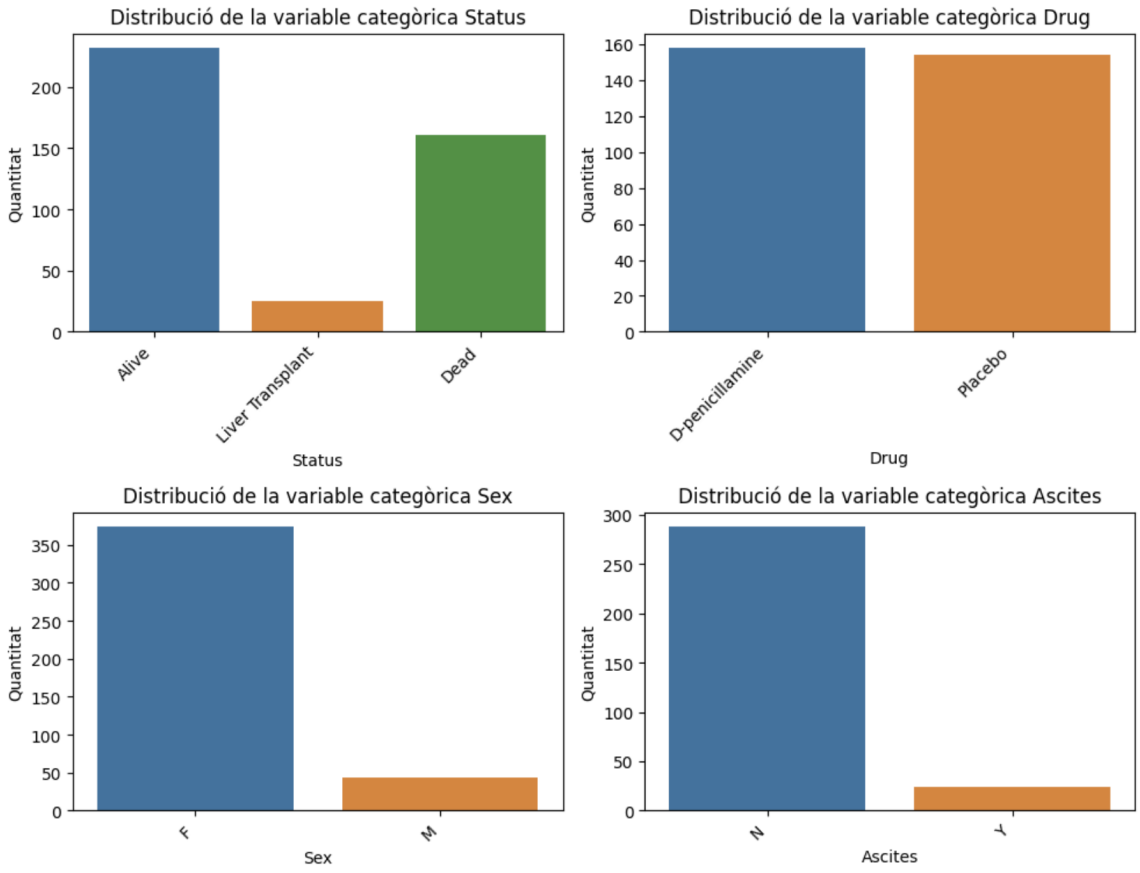

Figura 3: Countplots de variables categòriques del datset.

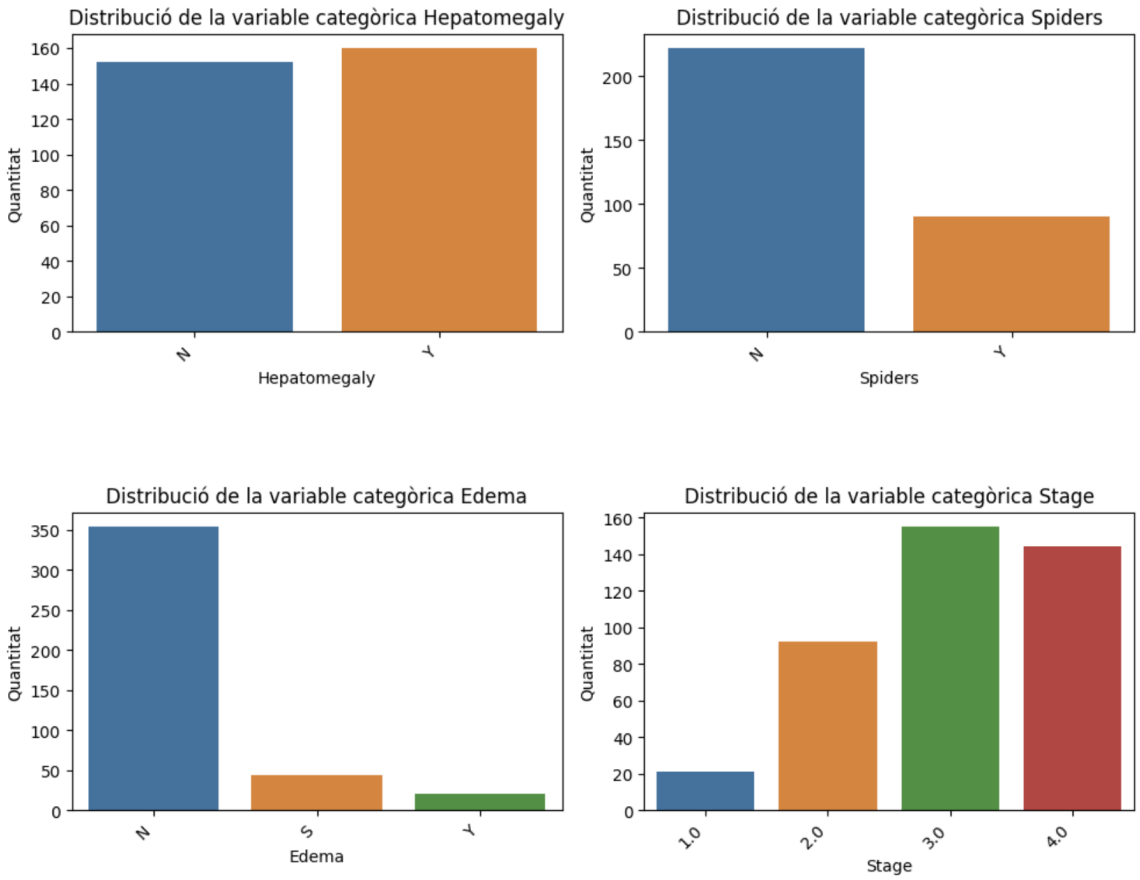

Figura 4: Countplots de variables categòriques del datset.

Es pot veure que les variables *Status* (la variable que determinarem com a *target* més endavant), *Sex*, *Ascites*, *Spiders*, *Edema* i *Stage* pateixen un clar desbalanceig de classes. Això s'haurà de tenir molt en compte a l'hora de

3.3 Missings

3.4 Outliers

3.5 Recodificació de variables

3.6 Particionat del dataset

4 Preparació de variables

4.1 Normalització de variables

4.2 Anàlisi de correlacions entre variables numèriques

4.3 Anàlisi de variables categòriques i variable objectiu

4.4 Eliminació de variables

4.5 Estudi de dimensionalitat (PCA)

5 Definició de models

5.1 K-Nearest Neighbors (KNN)

5.1.1 Motivació

5.1.2 Mètriques

5.1.3 Hiperparàmetres

5.1.4 Entrenament

5.1.5 Resultats

5.2 Arbre de decisió

5.2.1 Motivació

5.2.2 Mètriques

5.2.3 Hiperparàmetres

5.2.4 Entrenament

5.2.5 Resultats

5.3 Support Vector Machine (SVM)

5.3.1 Motivació

5.3.2 Mètriques

5.3.3 Hiperparàmetres

5.3.4 Entrenament

5.3.5 Resultats

6 Selecció del model

6.1 Descripció del model triat

6.2 Anàlisi de les limitacions i capacitats del model

6.3 Resultats

7 Model card

8 Bonus 1: Model EBM i comparació amb els models anteriors

9 Bonus 2: Anàlisi no supervisat de les dades

10 Conclusions

10.1 Valoració de l'aprenentatge adquirit

11 Referències

[1] E. Dickson; P. Grambsch; T. Fleming; L. Fisher; A. Langworthy. Cirrhosis Patient Survival Prediction. UCI Machine Learning Repository, 2023. DOI: https://doi.org/10.24432/C5R02G.