Informe sobre Chirrhosis Patient Survival Prediction dataset

Roger Baiges Trilla

28 de desembre de 2023

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

Abstract

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

A continuació s'esmenta el plantjeament realitzat per tal de resoldre el problema, els passos que s'han seguit i la implementació del codi en Python utilitzant llibreries com ara *pandas*, *numpy*, *sklearn*...

Finalment es comentaran els resultats obtinguts i les conclusions alhora que l'aplicativitat dels models en la realitat.

Contents

1 Introducció

El projecte es centra en el desenvolupament d'un model de *machine learning* per a la predicció de la supervivència de pacients amb cirrosi hepàtica. La cirrosi és una malaltia que resulta de la cicatrització del fetge, sovint causada per dany hepàtic a conseqüència de condicions com l'hepatitis B o C o l'ús crònic d'alcohol. Aquest dany no és reversible, però el tractament pot alentir el progrés de la malaltia, alleujar símptomes i prevenir complicacions. En les etapes inicials, és possible minimitzar el dany al fetge abordant les causes subjaents, com el tractament de l'addicció a l'alcohol, la pèrdua de pes, o l'ús de medicaments per tractar l'hepatitis viral i altres condicions [1].

El projecte utilitza un conjunt de dades amb 418 instàncies, recopilades en un entorn clínic, per construir models predictius. Aquestes dades contenen 17 característiques clíniques, cadascuna aportant informació vital sobre l'estat de salut dels pacients. Alguns dels símptomes freqüents de la cirrosi inclouen fatiga, pèrdua de pes, nàusees, fàcil aparició de blaus i sangrats, icterícia (coloració groguenca de la pell i els ulls), i acumulació de líquid a l'abdomen [2].

Des del punt de vista mèdic, aquestes dades ofereixen una oportunitat única per analitzar patrons i identificar factors predictius de la progressió de la malaltia i la supervivència dels pacients. Però un repte important en el treball amb aquestes dades és la gestió de valors perduts, particularment en les últimes files del conjunt de dades, on alguns pacients no van participar completament en l'experiment, encara que es van recollir algunes constants vitals.

L'objectiu del projecte és doble: primer, desenvolupar un model predictiu robust i fiable per a la supervivència dels pacients amb cirrosi, utilitzant tècniques avançades de *machine learning*; i segon, aportar comprensió mèdica més profunda de la cirrosi, una malaltia amb un gran impacte en la salut pública i en la gestió dels recursos sanitaris. La combinació de l'anàlisi estadística i el *machine learning* en aquest context mèdic no només millora la presa de decisions clíniques, sinó que també obre noves vies per a la investigació mèdica i la medicina personalitzada. Per tal de poder desenvolupar un bon treball ens farà falta saber primer a que ens estem enfrontant i, per tant, necessitem fer recerca científica de què tracten les variables.

El dataset utilitzat prové d'un estudi de la Mayo Clinic sobre la cirrosi biliar primària (PBC) del fetge realitzat entre 1974 i 1984, i inclou 17 característiques clíniques [3]. Aquestes variables són crucials per entendre la progressió de la malaltia i el pronòstic dels pacients. Algunes de les variables més rellevants són:

1. **Bilirubina**: Un producte de residu de la sang que passa a través del fetge i s'excreta en l'heces. Nivells elevats poden indicar dany hepàtic o malaltia [4].

2. **Albumina i proteïna total**: La albumina, una de les diverses proteïnes produïdes al fetge, és essencial per a la lluita contra infeccions i altres funcions. Nivells baixos poden ser indicatius de dany hepàtic o malaltia [4].

3. **Prothrombin time (PT)**: Mesura el temps que tarda la sang en coagular-se. Els increments en el PT poden indicar dany hepàtic, encara que també pot ser elevat per l'ús de medicaments anticoagulants com la warfarina [5].

4. **ALT i AST**: Enzims que, quan són alliberats en quantitats elevades a la sang, poden in-

dicar dany hepàtic o malaltia. L'ALT ajuda a convertir les proteïnes en energia per les cèl·lules hepàtiques, mentre que l'AST ajuda a descompondre els aminoàcids [4].

Aquestes variables, juntament amb altres com l'edat, el sexe, la presència d'ascites, hepatomegàlia, aranyes vasculars, edema, i els nivells de colesterol, entre d'altres, proporcionen una visió integral dels factors que afecten la supervivència dels pacients amb cirrosi [6]. El projecte pretén analitzar aquestes dades per desenvolupar models predictius que millorin la presa de decisions clíniques i contribueixin a la medicina basada en dades.

Els coneixement mèdics de la malaltia resultarn molt útils, per exemple, a l'hora de tractar amb els outliers on encara que un valor estigui molt allunyat de la mitjana pels humans, podria resultar en un valor comú per una persona amb cirrosi i, per tant, tot i ser un valor extrem l'haurem de considerar de forma diferent a que si fos un error.

2 Anàlisi i preprocessat de dades

En aquesta part del document s'analitzarà la base de dades fent un preprocessament juntament amb un estudi sobre les distribucions de les variables numèriques, els outliers i el valors faltants o missings. També es comentarà el balanceig de les dades per tal de si és necessari fer alguns ajustaments alhora d'entrenar models.

Per començar mitjançant la funció `data.shape` es va veure que contenia 418 files i 20 columnes. Posteriorment mitjançant la funció `data.describe()` es van obtenir les següents taules:

	ID	N_Days	Age	Bilirubin	Cholesterol	Albumin
count	418.00	418.00	418.00	284.00	418.00	
mean	209.50	1917.78	18533.85	3.22	369.51	3.49
std	120.81	1104.67	3815.85	4.41	231.94	0.42
min	1.00	41.00	9598.00	0.30	120.00	1.96
25%	105.25	1092.75	15644.50	0.80	249.50	3.24
50%	209.50	1730.00	18628.00	1.40	309.50	3.53
75%	313.75	2613.50	21272.50	3.40	400.00	3.77
max	418.00	4795.00	28650.00	28.00	1775.00	4.64

Table 1: Summary statistics for the numeric variables I

	Copper	Alk_Phos	SGOT	Triglycerides	Platelets	Prothrombin
count	310.00	312.00	312.00	282.00	407.00	416.00
mean	97.65	1982.66	122.56	124.70	257.03	10.73
std	85.61	2140.39	56.70	65.15	98.33	1.02
min	4.00	289.00	26.35	33.00	62.00	9.00
25%	41.25	871.50	80.60	84.25	188.50	10.00
50%	73.00	1259.00	114.70	108.00	251.00	10.60
75%	123.00	1980.00	151.90	151.00	318.00	11.10
max	588.00	13862.40	457.25	598.00	721.00	18.00

Table 2: Summary statistics for the numeric variables II

A continuació es presenta una interpretació estadística de les variables numèriques:

- **N_Days**: El nombre de dies de seguiment per als subjectes de l'estudi. La mitjana és de 1917.78 dies amb una desviació estàndard significativa de 3815.45 dies, indicant una variabilitat alta en el temps de seguiment entre els subjectes. El rang de la variable va des de 41 dies fins a un màxim de 28650 dies, suggerint la presència de valors extrems o outliers.
- **Age**: L'edat dels subjectes, expressada en dies. La mitjana és de 18533.85 dies, que equival a aproximadament 50 anys (assumint 365.25 dies per any per incloure anys de traspàs). Això indica que la població estudiada és de mitjana d'edat avançada. La desviació estàndard és de 3815.45 dies, mostrant també aquí una àmplia dispersió en les edats.
- **Bilirubin**: Els nivells de bilirubina en mg/dL. La distribució mostra una mitjana de 3.22 mg/dL, amb una variabilitat mesurada per una desviació estàndard de 4.41 mg/dL. Els valors

van des d'un mínim de 0.3 mg/dL fins a un màxim de 28 mg/dL, indicant que hi podrien haver subjectes amb nivells de bilirubina significativament elevats.

- **Cholesterol**: Els nivells de colesterol en mg/dL mostren una mitjana de 369.51 mg/dL. La desviació estàndard és de 231.94 mg/dL, el que indica una àmplia variabilitat en els nivells de colesterol entre els subjectes. El valor màxim registrat és de 1775 mg/dL, que està molt per sobre del rang considerat normal.
- **Albumin**: L'albumina, un important indicador de la funció hepàtica, mostra una mitjana de 3.49 g/dL, amb una desviació estàndard relativament baixa de 0.42 g/dL. Això suggereix que la majoria dels subjectes tenen nivells d'albumina dins d'un rang estret, proper al normal.
- **Copper**: Els nivells de coure en mcg/dL tenen una mitjana de 97.65 mcg/dL i una desviació estàndard de 85.61 mcg/dL, mostrant una variabilitat substancial entre els individus de l'estudi.
- **Alk_Phos**: La fosfatasa alcalina (Alk_Phos) en IU/L mostra una gran variabilitat amb una desviació estàndard de 2140.39 IU/L, molt elevada en comparació amb la mitjana de 1982.66 IU/L, suggerint la presència de valors atípics.
- **SGOT**: Una enzima hepàtica amb una mitjana de 122.56 U/L i una desviació estàndard de 56.70 U/L, indicant variabilitat moderada entre els subjectes.
- **Triglycerides**: Els nivells de triglicèrids en mg/dL mostren una mitjana de 124.70 mg/dL amb una desviació estàndard de 65.15 mg/dL. El rang és àmpli, amb valors que van des de 33 mg/dL fins a 598 mg/dL.
- **Platelets**: El nombre de plaquetes per microlitre de sang mostra una mitjana de 257.02 x $10^9/L$, amb una desviació estàndard de 98.33 x $10^9/L$. Això pot indicar una variabilitat significativa en la compta de plaquetes entre els pacients.
- **Prothrombin**: El percentatge de temps de protrombina, amb una mitjana de 10.73%, i una desviació estàndard de 1.02%, indica que la majoria dels subjectes tenen un temps de protrombina dins d'un rang relativament estret.

2.1 Preprocessat inicial

Inicialment al descarregar les dades originals mitjançant la llibreria *ucimlrepo* les dades venien sense preprocessar i es van realizar diverses modificacions per tal de que el propi *pandas* interpretés correctament les dades.

Les millores inicials que es van realitzar van ser les següents:

- Els valors en blanc o amb 'NANN' substituir-los per un missing data interpretable per *numpy* i *pandas*.
- Modificar els valores de la variable objectiu, *Status* ('C', 'D', 'LT') per uns més fàcils de llegir ('Alive', 'Death', 'Liver Transplant').
- Transformar totes les variables numèriques a *int64* o *float64* i les categòriques per *category*. Això és degut a que inicialment moltes variables eren considerades d'un tipus diferents a les seves dades.

Concretament les variables *Cholesterol*, *Tryglicerides*, *Copper*, *Platelets*, *Prothrombin*, *Alk_Phos*, *SGOT* i *Albumin* van necessitar canviar el seu tipus per tal de que *pandas* les interpretés com a valors enters o valors de punt flotant. Pel que fa a les categòriques tan sols la variables *Stage* es va canviar com a *category*.

Finalment per tal d'acabar aquest preprocessing es van comprobar que no hi haguessin files repetides en el dataset ja que això podria indicar que les dades estiguessin malamanet o inclús podria ocasionar problemes a l'entrenar models sobretot si algunes mostres es repetissin tant en el **train** com en el **test**.

2.2 Anàlisi estadístic

En aquesta secció s'estudiarà estadísticament com les dades es troben distribuïdes alhora que un breu estudi científic per tal d'analitzar si hi ha errors en elles o simplement els outliers es poden considerar valors anormals però possibles mèdicament sobretot per a persones amb malalties del fetge.

2.2.1 Anàlisi de la distribució de les variables

Per tal de poder analitzar les distribucions de les variables numèriques es van crear un histograma per a cada variable.

A més a més per tal de poder validar les conclusions extretes de les imatges es va crear una funció anomenada `evaluate_distribution()` que al rebre el nom de les columnes com a paràmetres anava una per una aplicant diferents test estadístics per tal de probar la distribució d'una variable. Per fer-ho es van agafar les principals distribucions:

- Normal
- Exponencial
- Log-normal
- Poisson

La funció calculava els *p-values* de cadascun d'aquests test i finalment retornava la distribució que seguia la variable en el cas que en que es pogués representar en alguna de les anteriors. També tenia en compte la *Skewness* i la *Kurtosis*, els quals serveixen per poder mesurar el grau de la desviació de la simetria i com de pronunciades són les cues comparades amb una normal, respectivament.

Pel que fa a la *Skewness* si el seu valor és positiu vol dir que la cua s'extén més cap a la dreta i, per tant, amb una concentració de dades superior a l'esquerra. D'altra banda si és negatiu la cua s'extén més cap a l'esquerra amb una concentració a la dreta.

La *Kurtosis* alta (leptocúrtica) ens indica que la distribució té cues més pesades i per tant més probabilitat per a valors extrems. En el cas que sigui baixa (platicúrtica) mostres que les cues són més lleugeres amb un pic més clar resultant amb menys probabilitat per als valors extrems.

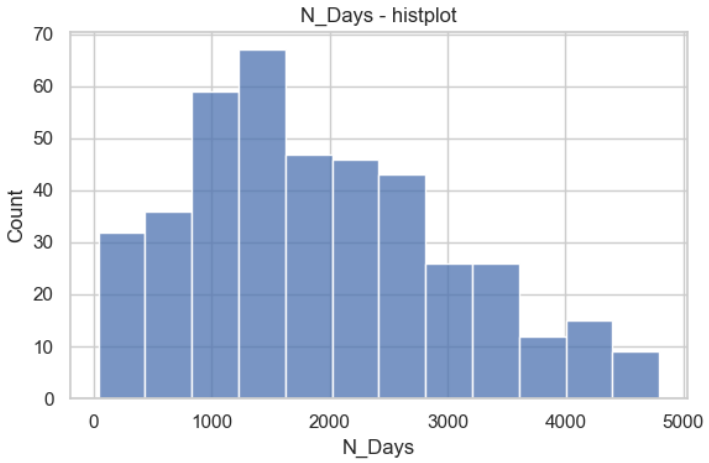

Figure 1: Imatge de la distribució de la variable N_Days

2.2.1.1 Variable N_Days La variable *N_Days* no s'ajusta a cap de les distribucions però degut als seu valor de 0.47 de *skewness* podem dir que té una cua a la dreta i que la majoria dels seus valors es centren a l'esquerra. Al mirar la *kurtosis* de -0.49 podem veure que és més plana que una distribució normal.

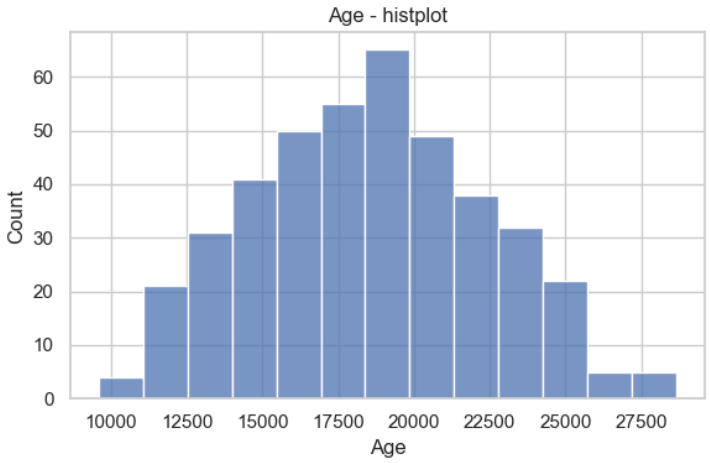

Figure 2: Imatge de la distribució de la variable Age

2.2.1.2 Variable Age La variable *Age* encara que tingui una forma similar a una normal estadísticament no ho és sinó que és una Log-Normal. Era d'esperar ja que la *skewness* és de 0.09, valor molt proper a 0, ja que es realitza comparant una distribució normal. D'altra banda la *kur-*

tosis de -0.62 significa que la gràfica és més plana que una normal.

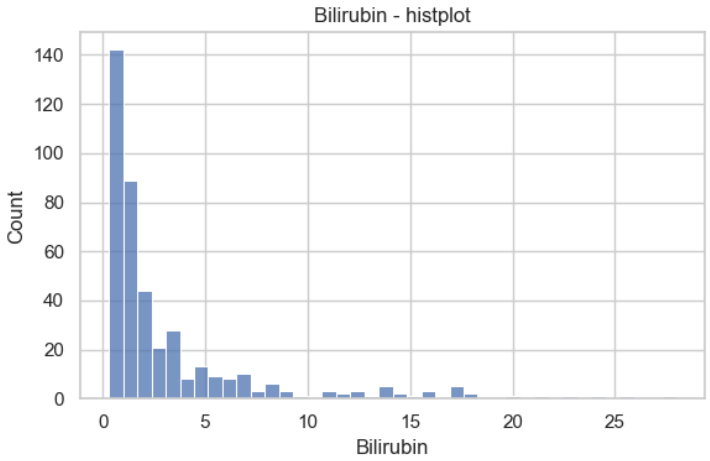

Figure 3: Imatge de la distribució de la variable Bilirubin

2.2.1.3 Variable Bilirubin La variable *Bilirubin* no s'adapta a cap de les distribucions. A més a més a simple vista destaca la gran cua que té a la dreta indicant molts possibles outliers. Al mirar la *skewness* veiem un valor de 2.71 que indica un gran biaix cap a la dreta com ja havíem vist. Cal destacar el valor de *kurtosis* de 7.95 que indica que la cua és molt més pesada que una normal.

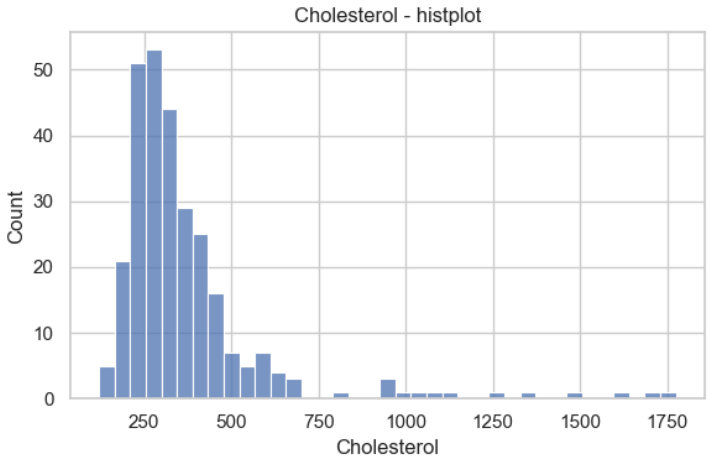

Figure 4: Imatge de la distribució de la variable Cholesterol

2.2.1.4 Variable Cholesterol La variable *Cholesterol* no s'ajusta a cap de les distribucions. Com es podia veure també en la figura 3 aquesta variable té una gran cua a la dreta indicant valors molt extrems. Amb uns valors de *skewness* i *kurtosis* de 3.39 i 14.07 respectivament veiem uns valors encara més exagerats que en el cas anterior. En la part dels outliers s'haurà de considerar el tractament d'aquesta variable.

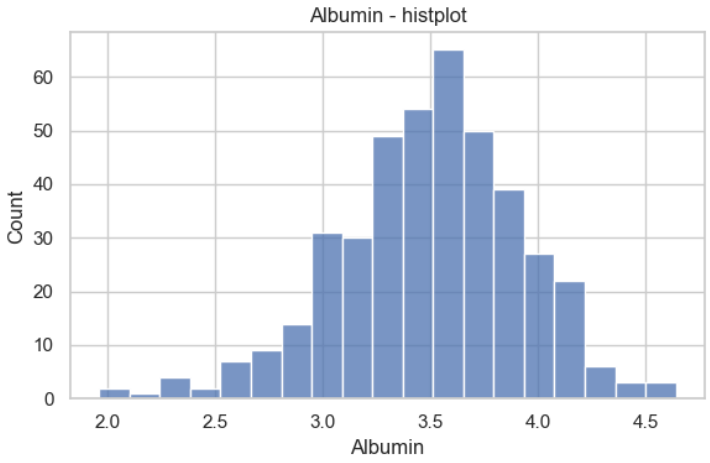

Figure 5: Imatge de la distribució de la variable Albumin

2.2.1.5 Variable Albumin La variable *Albumin* encara que a primera vista podría semblar una distribució normal, estadísticament no s'adapta a cap distribució. Els valors de *skewness* i *kurtosis* de -0.47 i 0.55 respectivament denoten que la variables té una lleugera cua a l'esquerra i que els valors es solen centrar més a la banda dreta tot i que no sembla que hi hagi outliers.

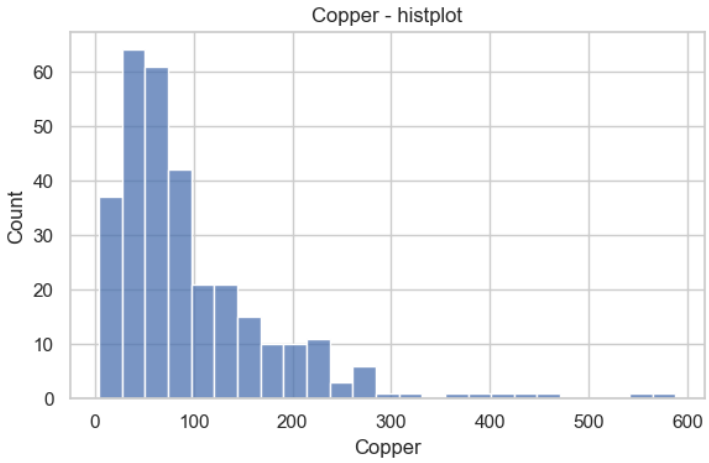

Figure 6: Imatge de la distribució de la variable Copper

2.2.1.6 Variable Copper La variable *Copper* sorprenentment s'adapta a una distribució Log-Normal amb una gran cua a la dreta que es veu reflexada en una *skewness* de 2.29 i una *kurtosis* de 7.48 indicant que és molt pesada.

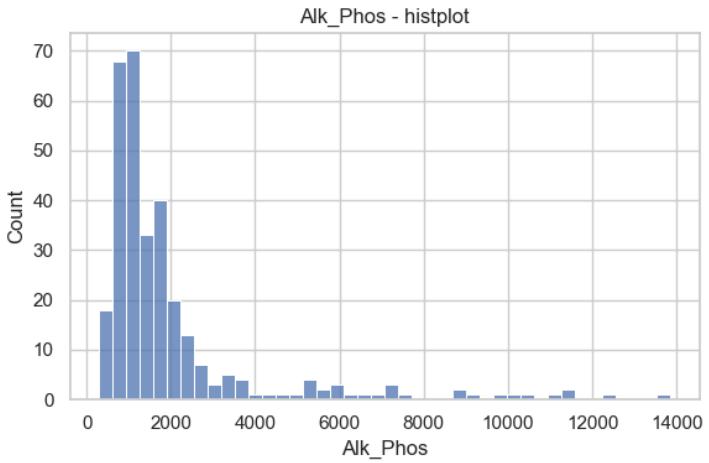

Figure 7: Imatge de la distribució de la variable Alk_Phos

2.2.1.7 Variable Alk_Phos La variable *Alk_Phos* no s'identifica amb cap de les distribucions i presenta una gran cua a la dreta. Els valors de *skewness* de 2.98 i *kurtosis* de 9.49 indiquen que hi ha extrems molt llunyans a una variable normal i que, molt probablement, els detectarem com a outliers.

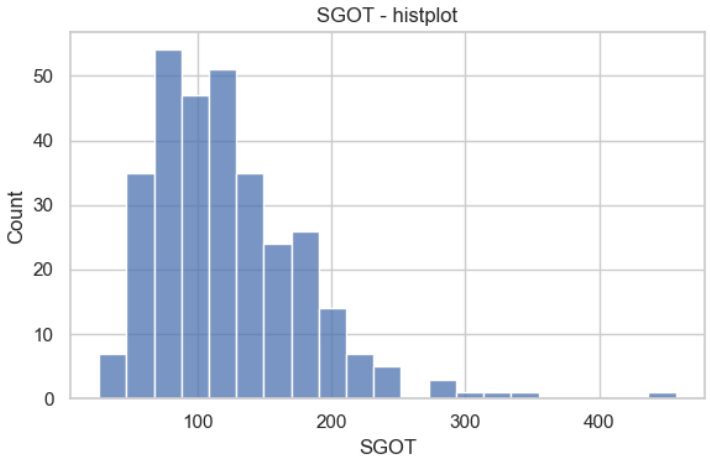

Figure 8: Imatge de la distribució de la variable SGOT

2.2.1.8 Variable SGOT La variable *SGOT* és considera una varialbe Log-Normal amb una cua també a la dreta. Els seus valors, per tant, es troben concentrats a l'esquerra amb una *skewness* i *kurtosis* de 1.44 i 4.22 respectivament.

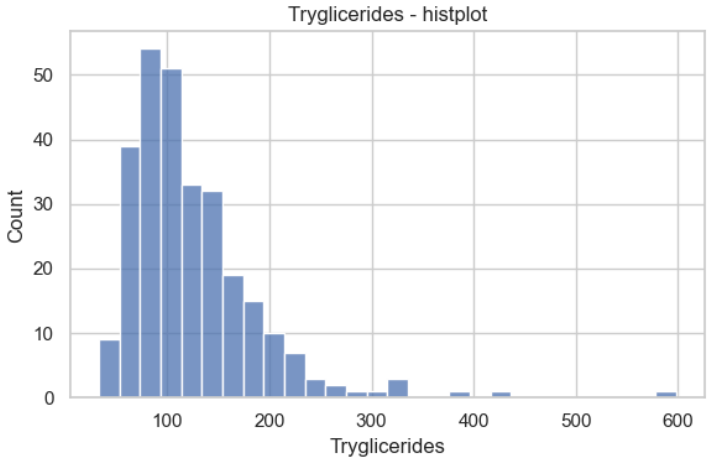

Figure 9: Imatge de la distribució de la variable Tryglicerides

2.2.1.9 Variable Triglycerides La variable *Tryglicerides* també segueix una distribució Log-Normal i també té una cua a la dreta. La seva distribució és molt semblant a la variable *SGOT* tot i que encara té més outliers degut al *skewness* de 2.51 i *kurtosis* d'11.57 significant un gran pes en aquesta cua.

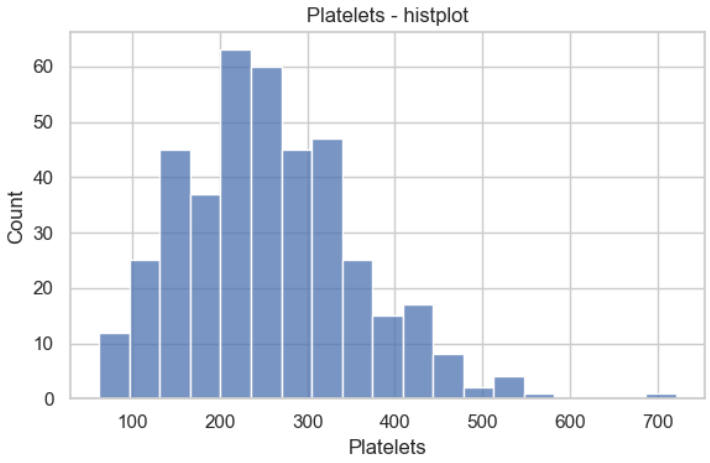

Figure 10: Imatge de la distribució de la variable Platelets

2.2.1.10 Variable Platelets La variable *Platelets* segueix una distribució Log-Normal desbiaixada lleugerament cap a l'esquerra amb un *skewness* de 0.62 i una *kurtosis* de 0.84 indicant que no s'allunya massa d'una normal.

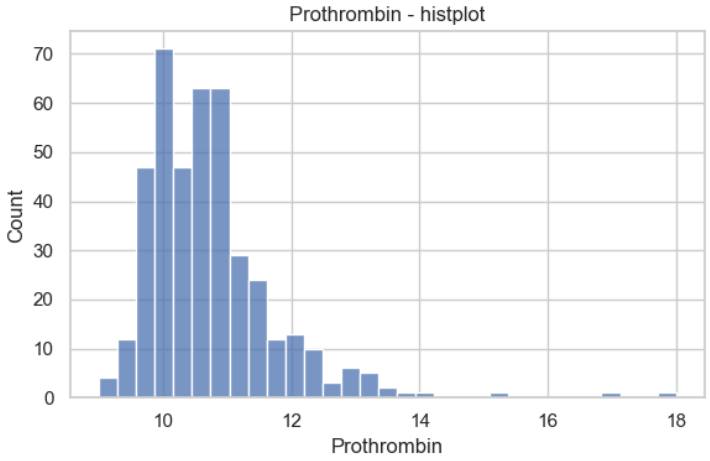

Figure 11: Imatge de la distribució de la variable Prothrombin

2.2.1.11 Variable Prothrombin La variable *Prothrombin* no pertany a cap distribució tot i que podem mencionar que té una gran cua a la dreta amb uns valors de *skewness* de 2.22 i de *kurtosis* de 9.92 indicant que hi ha outliers molt allunyats de la mitjana.

2.3 Estudi del balanceig de les dades

El conjunt de dades presenta diverses variables categòriques amb distribucions que mereixen una atenció especial, especialment pel que fa al desbalanceig de les classes. A continuació, es comenten algunes d'aquestes variables:

- **Status**: La variable objectiu mostra un desbalanceig significatiu amb 232 casos etiquetats com a 'Alive', 161 com a 'Dead', i només 25 com a 'Liver Transplant'. Aquest desbalanceig pot afectar la capacitat del model de machine learning per predir correctament els casos menys representats, com els de 'Liver Transplant'. Caldrà considerar-se com tractar aquest problema alhora d'entrenar els models. A més a més, faria falta informació sobre a qui feien trasplant de fetge degut a que es podria basar simplement en un factor sort i, per tant, els models no es poguessin acabar d'adaptar o simplement depenent d'algun patró com el nombre de dies que porten en l'experiment els pacients tenen més possibilitats de rebre'n un.

- **Drug**: La distribució entre les classes 'D-penicillamine' (158) i 'Placebo' (154) és relativament equilibrada, el que facilita l'anàlisi en aquest aspecte.

- **Sex**: Hi ha un clar desbalanceig de gènere, amb 374 casos de dones (F) i només 44 d'homes (M). Això pot introduir un biaix en el model, especialment si el gènere és un factor rellevant en la progressió de la malaltia. Com en el cas de la variable *Status* caldrà considerar-se a l'hora d'entrenar els models.

- **Ascites**, **Hepatomegaly**, **Spiders**, i **Edema**: Aquestes variables també mostren desbalanceig de classes, on la majoria de pacients no presenten aquests símptomes (com en el cas d'Ascites i Edema), excepte en Hepatomegaly on la distribució és més equilibrada.

- **Stage**: La distribució entre les etapes de la malaltia és relativament equilibrada entre les etapes 3 i 4, però hi ha menys representació en les etapes 1 i 2 indicant, per tant, que la majoria dels pacients de l'experiment tenien la malalta ja bastant desenvolupada.

A continuació, es presenta una taula amb el recompte dels valors de cada classe per a cada variable:

Variable	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**
Status	Alive (232)	Dead (161)	Liver Transplant (25)	-
Drug	D-penicillamine (158)	Placebo (154)	-	-
Sex	F (374)	M (44)	-	-
Ascites	N (288)	Y (24)	-	-
Hepatomegaly	Y (160)	N (152)	-	-
Spiders	N (222)	Y (90)	-	-
Edema	N (354)	S (44)	Y (20)	-
Stage	1 (21)	2 (92)	3 (155)	4 (144)

Table 3: Recompte de Valors de les Classes per a cada variable Categòrica

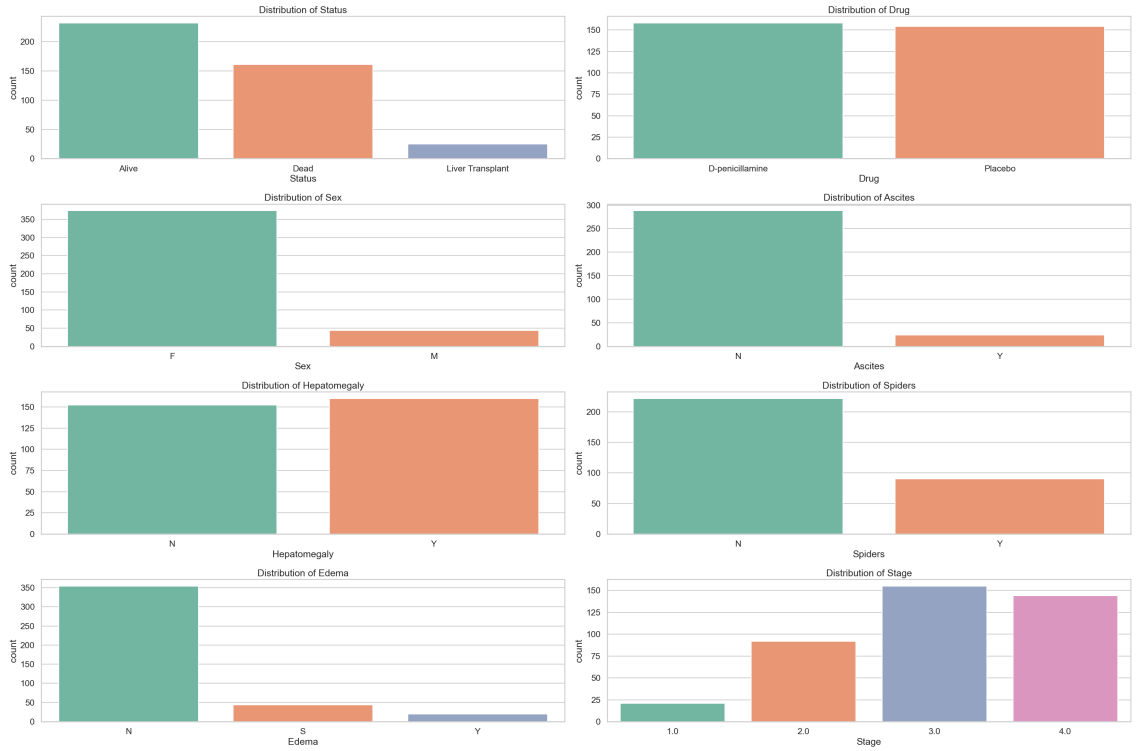

Figure 12: Imatge dels barplots de les variables categòriques

2.4 Estudi dels missings

L'existència de valors faltants en les dades pot representar un desafiament significatiu en el desenvolupament de models de *machine learning*, ja que pot afectar la qualitat de les prediccions i requereix una atenció especial en el preprocessament de les dades. A continuació es presenta una anàlisi dels valors perduts per cada variable:

- Les variables **Triglicèrids** i **Colesterol** presenten els majors percentatges de valors perduts, amb un 32.54% i 32.06% respectivament, el que pot dificultar l'anàlisi d'aquests factors relacionats amb el metabolisme hepàtic.
- **Coure**, **Fàrmac**, **Ascitis**, **Hepatomegàlia**, **Aranyes**, **Alk_Phos** i **SGOT** mostren un 25.36% de valors perduts, indicant que una quarta part de les dades estan incompletes per aquestes variables clíniques importants.
- La variable **Plaquetes** i **Etapa** tenen pèrdues molt més baixes, amb un 2.63% i 1.44% respectivament, el que indica una major integritat de les dades en aquests camps.
- **Prothrombin** mostra un percentatge mínim de valors perduts (0.48%), suggerint que la majoria de les dades de temps de protrombina estan disponibles per a l'anàlisi.

La gestió d'aquests valors perduts pot incloure tècniques com la imputació, l'eliminació de files, o l'ús de models que puguin manejar dades incompletes. La tria de la tècnica de gestió de valors

perduts ha de ser informada per l'anàlisi exploratòria de les dades i les consideracions específiques del domini de la malaltia.
La taula següent resumeix els valors perduts per a cada variable:

Variable	Total Missing	Percentage
Triglycerides	136	32.54%
Cholesterol	134	32.06%
Copper	108	25.84%
Drug	106	25.36%
Ascites	106	25.36%
Hepatomegaly	106	25.36%
Spiders	106	25.36%
Alk_Phos	106	25.36%
SGOT	106	25.36%
Platelets	11	2.63%
Stage	6	1.44%
Prothrombin	2	0.48%

Table 4: Missing data per Variable

2.4.1 Funció per imputar el missing data

Per tal de poder tractar amb aquest valors faltants es va crear una funció degut a que les dades s'haurien d'imputar diverses vegades al llarg dels diferents entrenaments dels models com ara en el dataset del **train**, totes les particions del **validation data** i finalment en el **test**.

La funció `impute_dataset` s'encarrega d'imputar valors perduts dins d'un conjunt de dades. La funció treballa separadament amb variables numèriques i categòriques:

- Per a les **variables numèriques**, s'utilitza `KNNImputer`, una tècnica d'imputació basada en els k veïns més propers. Aquesta elecció es deu a la seva eficàcia en capturar la estructura subjaent de les dades, permetent imputar valors de manera informada a partir del context proporcionat pels veïns més propers.
- Per a les **variables categòriques**, es crea un model `RandomForestClassifier` per a cada columna categòrica. Aquest enfocament es justifica pel fet que el *Random Forest* pot modelar la no linealitat i la complexitat de les relacions entre les variables. S'entrena únicament amb variables numèriques per evitar la complexitat addicional que implica l'*encoding* de les variables categòriques i perquè la major presència de variables numèriques en el conjunt de dades fa que els resultats d'imputació no variïn substancialment al incloure les variables categòriques.

La funció comença amb una còpia del conjunt de dades original i segueix aquests passos:

1. Identifica les columnes numèriques i categòriques.
2. Normalitza, per defecte està en **True**, les variables numèriques mitjançant `Standard Scaler`.
3. Aplica `KNNImputer` a les columnes numèriques.

4. Transforma de nou als valor numèrics reals, en el cas que s'hagués aplicat la normalització.
5. Per cada columna categòrica amb valors perduts, entrena un `RandomForestClassifier` amb les dades on la columna està completa, usant la resta de columnes numèriques com a predictors.
6. Imputa els valors perduts en la columna categòrica utilitzant les prediccions del model.

Aquest enfocament assegura que l'imputació es realitza de manera específica per al tipus de dades, preservant la integritat i la relació entre les variables. La qualitat de la imputació s'avaluarà en el següent pas de l'anàlisi, on s'inspeccionaran els resultats i es compararan amb les característiques originals de les dades.

2.4.2 Funció per evaluar la qualitat de la imputació

Per validar la qualitat de la imputació realitzada per la funció `impute_dataset`, es fan servir diverses mètriques específiques per a dades numèriques i categòriques:

- Per a dades **numèriques**, s'empren mètriques com l'error quadràtic mitjà (*Mean Squared Error*, MSE), el coeficient de determinació (R^2), l'error absolut mitjà (*Mean Absolute Error*, MAE), l'error quadràtic mitjà arrel (RMSE), i l'error percentual absolut mitjà (MAPE). Aquestes mètriques proporcionen una visió completa de la precisió i la qualitat de la imputació.
- Per a dades **categòriques**, s'utilitzen l'exactitud (*Accuracy*) i la puntuació F1 (*F1 Score*). L'exactitud mesura la proporció de valors correctament imputats, mentre que la puntuació F1 proporciona una mesura harmònica de la precisió i la recuperació.

Les funcions `hide_data` i `evaluate_imputation` es fan servir per amagar i després validar la imputació. Aquestes funcions són claus per simular la pèrdua de dades i per avaluacions posteriors:

- La funció `hide_data` oculta una fracció de les dades originals d'una columna determinada, permetent simular la pèrdua de dades i validar la capacitat de la funció d'imputació per recuperar aquests valors.
- La funció `evaluate_imputation` fa servir la funció `hide_data` per amagar les dades, aplica la imputació, i després compara els valors imputats amb els originals per calcular les mètriques de validació. Això ofereix una avaluació directa de la imputació en un entorn controlat.

El procés de validació s'executa per totes les columnes del conjunt de dades, calculant les mètriques específiques per a cada tipus de dada i proporcionant un informe detallat de la qualitat de la imputació per a cada variable.
Les mètriques mitjanes calculades a partir de les mètriques individuals proporcionen una vista agregada de la qualitat de la imputació a través del conjunt de dades. Aquest enfocament integral assegura que l'imputació sigui validada de manera rigorosa i informada.
El codi següent mostra la implementació d'aquestes funcions i com s'utilitzen per calcular les mètriques:

```
# Function to randomly hide data in a column
def hide_data(df, column, hide_ratio=0.1):
    ...
```

```
# Function to evaluate imputation
def evaluate_imputation(df, column, hide_ratio=0.1):
    ...

# Calculate average metrics for all variables
...
```

Aquesta validació rigorosa és essencial per assegurar que la imputació sigui fiable i que els models de *machine learning* construïts a partir d'aquestes dades imputades siguin robustos i precisos.

2.4.3 Anàlisi dels resultats de la imputació

L'avaluació de la qualitat de la imputació s'ha realitzat utilitzant mètriques estàndard en la ciència de dades. Els resultats obtinguts mostren variabilitat en la qualitat de la imputació entre les variables numèriques i categòriques. A continuació, es presenten les taules amb els resultats obtinguts per a cada grup de variables.

Per poder obtindre els resultats s'ha imputat amb la llavor *42* el dataset original; això ens pot donar una idea dels resultats que s'obtindran a l'imputar els diferents dataset ja mencionats anteriorment.

Variables Numèriques La Taula 5 mostra els resultats de la imputació per a les variables numèriques. Els resultats indiquen que, per a algunes variables com `N_Days` i `Age`, els valors MSE i MAPE són relativament alts. Això pot ser atribuït a la presència d'outliers i l'ampli rang de valors dins del conjunt de dades. Per exemple, la variable `N_Days` presenta una gran variabilitat, amb alguns pacients tenint poques dies i altres molts més, el que pot fer que qualsevol error d'imputació tingui un impacte desproporcionat en les mètriques.

Variable	**MSE**	R^2	**MAE**	**RMSE**	**MAPE (%)**
N_Days	1466164.98	0.1991	916.37	1210.85	167.35
Age	15535161.20	0.0486	3357.62	3941.47	19.52
Bilirubin	9.29	0.4522	2.11	3.05	115.71
Cholesterol	86356.98	0.2142	163.20	293.87	36.62
Albumin	0.13	0.3203	0.31	0.36	9.30
Copper	4013.14	-0.0863	49.12	63.35	56.89
Alk_Phos	1222259.89	0.0013	842.86	1105.56	65.67
SGOT	5481.23	-0.0228	48.99	74.04	34.39
Tryglicerides	2917.34	0.1828	36.24	54.01	24.69
Platelets	9627.02	-0.2759	71.02	98.12	23.69
Prothrombin	0.48	0.0714	0.51	0.69	4.60
Mitjana	**1666544.70**	**0.1004**	**498.94**	**622.31**	**50.77**

Table 5: Resultats d'imputació per a variables numèriques

Variables Categòriques Per altra banda, les variables categòriques, com es mostra a la Taula 6, han obtingut resultats més positius en la imputació. Destaca la variable objectiu `Status`, amb una exactitud (*Accuracy*) de 0.78 i una puntuació F1 de 0.76, suggerint que l'imputació pot ser fiable i útil per a models posteriors. Cal destacar els resultats pràcticament aleatoris de la variable *Drug*;

això és degut a que aquesta variable realment no es pot predir degut a que no és un factor que una persona pateix o no sinó simplement té a veure en un factor sort on un metge decideix si un pacient li donaran la malaltia real o tan sols un *placebo*.

Variable	**Accuracy**	**F1 Score**
Status	0.7857	0.7665
Drug	0.4839	0.4728
Sex	0.8810	0.8252
Ascites	0.9355	0.9355
Hepatomegaly	0.7097	0.7116
Spiders	0.7419	0.7165
Edema	0.8571	0.8010
Stage	0.4390	0.4366
Mitjana	**0.7292**	**0.7082**

Table 6: Resultats d'Imputació per a Variables Categòriques

Els resultats de la imputació per a les variables categòriques amb una mitjana d'accuracy de 0.73 i un f1 de més de 0.7 suggereixen que els models de classificació utilitzats per a la imputació podrien ser efectius en predir la classe de les dades censurades o perdudes. Aquesta eficàcia és especialment notable en la variable objectiu, que és crítica per a la predicció de la supervivència en el conjunt de dades.
En resum, mentre que les variables numèriques poden requerir una atenció addicional en la gestió d'outliers per millorar la qualitat de la imputació, les variables categòriques mostren resultats prometedors que poden ser utilitzats en models predictius amb confiança.

2.5 Estudi dels outliers

L'anàlisi d'outliers és un pas crític en la preparació de dades per models de *machine learning*. L'existència d'outliers pot afectar significativament l'aprenentatge i la predicció dels models. Es presenten dues funcions per a identificar i eliminar aquests outliers:

- La funció `outliers_analysis` identifica els outliers utilitzant el rang interquartílic (IQR) per cada variable numèrica, definint els límits superiors i inferiors i comptant el nombre i percentatge d'outliers. Això ens pot permetre identificar quins valors són considerats extrems i pocs comuns tot i que per acabar de determinar-ho farà falta un estudi científic per confirmar-ho.
- La funció `remove_outliers` elimina els outliers basant-se en els límits calculats, amb l'opció de ser més o menys estrictes amb aquests límits mitjançant un multiplicador.

Les variables com `Bilirubin`, `Copper`, `Cholesterol`, i `Alk_Phos` són particularment susceptibles als outliers i requereixen un tractament específic. La Taula 7 mostra el resultat de l'anàlisi d'outliers per cada variable numèrica:

Variable	IQR	Lower Bound	Upper Bound	Outliers Count	Percentage (%)
N_Days	1520.75	-1188.38	4894.63	0	0.00
Age	5628.00	7202.50	29714.50	0	0.00
Bilirubin	2.60	-3.10	7.30	46	11.00
Cholesterol	150.50	23.75	625.75	20	4.78
Albumin	0.53	2.45	4.56	9	2.15
Copper	81.75	-81.38	245.63	17	4.07
Alk_Phos	1108.50	-791.25	3642.75	35	8.37
SGOT	71.30	-26.35	258.85	7	1.67
Tryglicerides	66.75	-15.88	251.13	10	2.39
Platelets	129.50	-5.75	512.25	6	1.44
Prothrombin	1.10	8.35	12.75	18	4.31

Table 7: Anàlisi d'Outliers per a Variables Numèriques

Com es pot veure en la taula 7 hi ha valors negatius en variables que no poden tenir valors negatius; això realment no comparta cap problema i simplement vol dir que degut que la majoria de dades es troben concentrades a l'esquerra, amb valors propers al 0, com ja s'ha vist en l'estudi de les distribucions amb la *skewness* en l'anàlisi de les distribucións de les variables.

Les variables `Bilirubin`, `Copper`, `Cholesterol` i `Alk_Phos` presenten un percentatge d'outliers superior al 4%, indicant la presència de valors atípics significatius en comparació amb altres variables. Aquesta observació justifica l'ús d'un multiplicador més baix en la funció `remove_outliers` per a ser més restrictius amb aquests límits quan es treballa amb aquestes variables. L'eliminació selectiva d'outliers pot ajudar a reduir l'impacte dels valors extrems i millorar la precisió dels models de predicció.

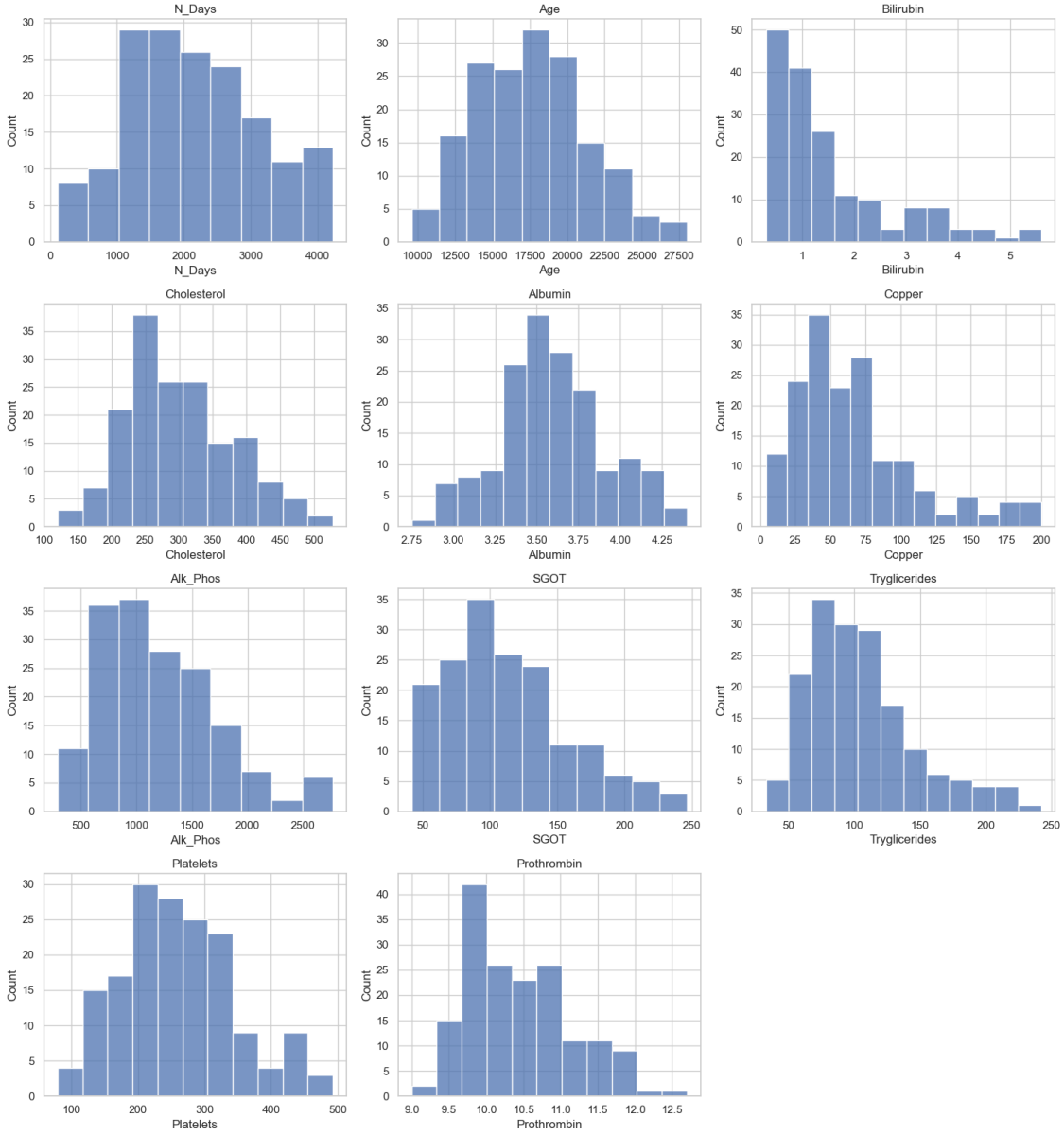

Figure 13: Imatge de les distribucions de les variables numèriques després d'eliminar els outliers

Com es pot observar en la figura 13 i si ho comparem amb les distribucions originals de 2.2.1 podem observar que les distribucions es troben lleugerament afectades. Destaquem grans canvis en les variables que ja havíem mencionat que tenien molts outliers.

2.6 Anàlisi mèdic dels outliers

2.6.1 Bilirubin

La bilirubina és un pigment groguenc trobat a la bilis, un líquid produït pel fetge. Un test de sang mesura els nivells de bilirubina, i nivells elevats poden indicar problemes de funció hepàtica [7].
Els nivells normals de bilirubina són [7]:

- Directa (també anomenada conjugada): menys de 0.3 mg/dL
- Total: de 0.1 a 1.2 mg/dL

Els nivells de bilirubina en pacients amb cirrosi poden variar i no hi ha un nivell fix de bilirubina que diagnostiqui la cirrosi [8]. Moltes persones amb cirrosi no presenten nivells elevats de bilirubina i, per tant, valors normals de bilirubina no exclouen la cirrosi. Diverses condicions poden afectar els nivells de bilirubina, incloent el Síndrome de Gilbert, l'hemòlisi (destrucció anormal de glòbuls vermells), i l'hepatitis infecciosa, entre d'altres [9].
Els nivells elevats de bilirubina en la cirrosi normalment indiquen una eliminació de bilirubina reduïda o una producció alta sense una eliminació suficient [10].
A continuació es presenta una taula amb els valors normals i els valors trobats en pacients amb cirrosi:

Tipus de Bilirubina	**Nivells Normals (mg/dL)**	**Nivells en Cirrosi (mg/dL)**
Directa	Menor de 0.3	Pot variar
Total	0.1 - 1.2	Pot ser elevat

Table 8: Comparativa de Nivells de Bilirubina

Saben aquesta informació podem arribar a la conclusió que és comú pels malalts de cirrosis tenir una bilirubina eleveda així que en el models s'haurà de comparar si els resultats milloren o no amb aquests outliers enlloc d'eliminar-los directament ja que mèdicament poden ser possibles.

2.6.2 Cholesterol

El colesterol és una substància cerosa trobada en totes les cèl·lules del cos i és essencial per a la creació de membranes cel·lulars, hormones i altres substàncies vitals. El cos necessita colesterol per a funcionar correctament, però nivells excessius poden augmentar el risc de malalties cardíaques [11].
Els nivells normals de colesterol són [12]:

- LDL (colesterol "dolent"): menys de 100 mg/dL
- HDL (colesterol "bo"): 40 mg/dL o més
- Total: menys de 200 mg/dL

En pacients amb cirrosi, els nivells de colesterol poden veure's afectats per diversos factors, incloent la mala absorció de greixos i una síntesi reduïda de colesterol en el fetge [13]. No és inusual trobar pacients amb cirrosi que tenen nivells baixos de colesterol total, sobretot si la funció hepàtica està compromesa significativament [14].

A continuació es presenta una taula amb els valors normals i els valors trobats en pacients amb cirrosi, en comparació amb les dades del dataset proporcionades:

Tipus de Colesterol	Nivells Normals (mg/dL)	Nivells en Cirrosi (mg/dL)
LDL	Menor de 100	Pot ser baix
HDL	40 o més	Pot variar
Total	Menys de 200	Pot ser baix

Table 9: Comparativa de Nivells de Colesterol

Amb aquesta informació, es pot inferir que els nivells baixos de colesterol en pacients amb cirrosi poden ser una manifestació de la malaltia. Per tant, en el context del modelatge, s'hauria de considerar aquests valors com a potencialment informatius i no necessàriament com a outliers que cal eliminar. La presència de valors atípics de colesterol podria reflectir la gravetat de la cirrosi i, per tant, hauria de ser analitzada en els propis models predictius de la malaltia.

En el dataset també es poden observar valors que podrien ser errors de mesura com ara el valor màxim de 1775 quan en un adult normal la mitjana és menor que 200 i en el cas de persones amb malalties hepàtiques també és considera massa elevat.

Cal mencionar també que en el dataset trobem pacients que es troben en diferents etapes de la malaltia cosa que pot afectar als valors fent que siguin més elevats del que haurien.

2.6.3 Copper

El coure és un mineral traça essencial per a la salut humana. Un excés de coure en el cos pot ser perjudicial i està associat amb condicions com la malaltia de Wilson i d'altres disfuncions hepàtiques [15].

Els nivells normals de coure en el sèrum solen estar entre 70 i 150 g/dL. Nivells elevats poden indicar una acumulació de coure deguda a una excreció inadequada, una situació possible en el context de la cirrosi hepàtica [16].

Com s'ha pogut observar en l'histograma 6, podem trobar les dades en els rangs normals i esperats ja que la majoria dels pacients presenten nivells baixos a moderats de coure, però hi ha una presència notable d'outliers amb valors molt elevats. Aquests casos podrien ser el resultat d'una malaltia hepàtica avançada o d'altres condicions mèdiques, i per tant, s'hauria d'avaluar la seva pertinença en el conjunt de dades abans de prendre decisions sobre el seu tractament en l'anàlisi de dades.

2.6.4 Alk_Phos

La fosfatasa alcalina és una enzim clau en els processos metabòlics del fetge i dels ossos. Nivells elevats en sang poden ser indicatius de malalties hepàtiques com la cirrosi, on l'enzim pot incrementar-se com a resposta a obstruccions biliars o a dany hepàtic [17].

Els nivells normals d'‘Alk_Phos‘ són generalment entre 44 i 147 U/L. Els pacients amb cirrosi poden presentar valors elevats, els quals han de ser avaluats en context clínic [18].

L'anàlisi del conjunt de dades mostra una distribució amb una majoria de valors d'*Alk_Phos* superiors als 1500 cosa que pot indicar que la unitat del metadata file és incorrecta. Degut a que la majoria de valors són tan grans faria falta consultar a un professional de la salut per acabar de confirmar els resultats obtinguts. Com que probablement es tracti d'una confusió analitzarem la variable normalment considerant outliers els que estiguin unes 3 o 4 vegades més lluny del Q3.

2.6.5 Prothrombin

El temps de protrombina és una mesura crítica de la capacitat de coagulació de la sang, i és especialment rellevant en el context de malalties hepàtiques com la cirrosi, on la síntesi de factors de coagulació pot estar compromesa [19].

Els nivells normals per al temps de protrombina solen estar al voltant de 11 a 13.5 segons, amb una relació normalitzada internacional (INR) d'aproximadament 0.8 - 1.1. Amb una mitjana de 10.7 segons i un desviament estàndard de 1.02 en el nostre conjunt de dades, la majoria dels pacients semblen tenir un temps de protrombina dins del rang normal. Aquesta observació pot indicar que, tot i la presència de cirrosi, la funció de coagulació encara es manté dins dels límits acceptables per a molts dels individus estudiats. Tanmateix, la presència d'una certa variabilitat dins dels pacients reflecteix la complexitat de la cirrosi i la seva afectació heterogènia en la coagulació sanguínia.

Serà important per a qualsevol model de machine learning considerar aquesta variabilitat i entendre el seu possible impacte en els resultats de salut dels pacients. Això pot requerir una anàlisi més granular de la relació entre el temps de protrombina i altres variables clíniques dins del conjunt de dades.

2.6.6 Variables Albumin, SGOT, Trygliceride i Platelets

Com es pot observar en la taula 10 en totes aquestes variables que contenen outliers els seus valors no estan tan allunyats dels valors considerats mèdicament normals per una persona adulta. L'única variable que destaca pels seus valors més elevats és $SGOT$ degut a que la seva mitjana és fins a 3 vegades els valores comuns per a persones sanes però això ja és esperat degut a que la cirrosi pot causar dany hepàtic, resultant en la lliberació d'enzims hepàtics com l'SGOT a la sang.

Biomarcador	Rang Normal	Mitjana en Cirrosi	Desviació Estàndard
Albumina (g/dL)	3.4 - 5.4	3.5	0.42
SGOT (U/L)	10 - 40	122	56.7
Triglicèrids (mg/dL)	150 - 200	124	65
Plaquetes (per microlitre)	150,000 - 450,000	257,000	98,000

Table 10: Comparació dels Biomarcadors en Pacients amb Cirrosi vs Valors Normals

2.6.7 Conclusions de l'anàlisi científic

Com a conclusions s'ha de dir que fent recerca a internet no és suficient per tal de poder treure conclusions sino que faria falta consultar amb un professional o expert.

Cal mencionar però que aquests resultats mèdics, en general, informen que les variables amb outliers realment són realistes i la majoria no es poden considerar errors si més no valors molt elevats probablement deguts a l'estat en la qual es troba el pacient en aquesta malaltia.

2.7 Recodificació de les variables

Pel que fa a la recodificació de variables això és un pas molt important i necessari alhora d'entrenar models ja que la majoria de models, sobretot els que es faran servir en el projecte, requereixen un dataset amb tan sols vairables numèriques.

Per tal d'aconseguir això el que cal fer és convertir les variables categòriques a numèriques mitjançant diferents tècniques d'*encoding* com ara `OneHotEncoding`, `OrdinalEncoding` o `LabelEncoding`.

El `OneHotEncoding` converteix cada valor únic d'una categoria en una nova columna i assigna un valor binari. És útil quan les categories no tenen un ordre o jerarquia natural. Un dels principals defectes que té és que augmenta molt la dimensionalitat de la base de dades cosa que pot complicar molt més els càlculs dels propis models.

El `OrdinalEncoding` assigna un valor únic a cada categoria en format d'un nombre sencer. Les categories s'ordenen i s'assigna un nombre basat en aquest ordre. No augmenta la dimensionalitat.

El `LabelEncoding` és similar a l'OrdinalEncoding encara que no assumeix un ordre jeràrquic. A vegades pot liar els models malinterpretant les dades.

Finalment la decisió de fer servir un tipus o un altre d'encoding dependrà del propi model ja que per exemple en models d'arbres com ara *DecisionTree* o *RandomForest*, és més útil utilitzar encodings que permeten un ordre per tal de poder fer comparacions. Per tant, la decisió de fer servir un tipus o un altre dependrà del propi model a entrenar.

2.8 Particionat del dataset

El dataset s'ha particionat en un 80% per a l'entrenament i un 20% per a la prova. Aquesta decisió es basa en la necessitat d'equilibrar entre tenir suficients dades per a l'entrenament del model i la capacitat de validar de manera efectiva el rendiment del model. Amb un dataset relativament petit de 418 files (això si no s'elimina cap degut als missing o outliers), reservar un 80% per a l'entrenament assegura que el model tingui prou exemples per aprendre de manera efectiva.

2.8.1 Ús de Cross-Validation

El mètode de validació creuada (cross-validation) s'aplica dins del conjunt del **train** per diverses raons clau:

1. **Optimització dels Hiperparàmetres:** Cross-validation permet una avaluació més precisa dels hiperparàmetres. En lloc de limitar-se a un únic conjunt de validació, el model es prova en múltiples subconjunts, proporcionant una millor estimació de com els hiperparàmetres funcionaran en dades no vistes.
2. **Maximització de les Dades d'Entrenament:** Amb un dataset de mida limitada, és crucial utilitzar les dades de la manera més eficient possible. Cross-validation assegura que totes les dades d'entrenament s'utilitzen tant per a l'aprenentatge com per a la validació en diferents iteracions, evitant la necessitat d'un conjunt de validació separat que reduiria la quantitat de dades disponibles per a l'entrenament.
3. **Reducció de la Variància:** El model es valida múltiples vegades en diferents conjunts d'entrenament i validació, reduint la variància dels resultats de la validació i proporcionant una millor generalització.

En conclusió, el particionat del dataset en un 80% per al **train** i un 20% per al **test**, juntament amb l'ús de **cross-validation**, és una estratègia eficaç per a optimitzar el rendiment del model en un dataset de mida tan reduïda.

2.9 Balancejar les dades

El balanceig de dades és una tècnica crucial per tractar amb datasets on les classes objectiu estan desbalancejades. Aquest desequilibri pot portar a un model d'aprenentatge automàtic a sobreestimar la importància de les classes majoritàries.

2.9.1 Funció de Balanceig de Dades

La funció `balance_dataset` proporciona una estratègia per tal d'eliminar el desequilibri en les classes d'un dataset de machine learning. Utilitza els mòduls `SMOTE` i `RandomUnderSampler` de la biblioteca `imbalanced-learn` per aplicar oversampling a les classes minoritàries i undersampling a les classes majoritàries.

2.9.1.1 Mecànica de la Funció

```
def balance_dataset(X, y, do_oversample=False, do_undersample=False,
                    random_state=42):
    # Contingut de la funció
```

La funció `balance_dataset` pren un conjunt de dades representat per X (atributs) i y (etiquetes de classe). Els paràmetres `do_oversample` i `do_undersample` determinen si s'aplicarà oversampling, undersampling, o ambdós.

2.9.1.2 Oversampling amb `SMOTE` Si `do_oversample` està activat, la funció utilitza `SMOTE` per generar mostres sintètiques de les classes minoritàries, intentant augmentar la seva representació al nivell de les classes majoritàries. Això pot ajudar a millorar el rendiment dels algoritmes de classificació que podrien estar biaixats cap a les classes majoritàries.

2.9.1.3 Undersampling amb `RandomUnderSampler` Si `do_undersample` està activat, la funció fa servir `RandomUnderSampler` per eliminar mostres de les classes majoritàries. Aquesta tècnica busca reduir la bretxa entre les classes majoritàries i minoritàries, però pot portar a una pèrdua d'informació valuosa.

2.9.1.4 Combinació d'Oversampling i Undersampling La funció pot aplicar una combinació d'oversampling i undersampling per equilibrar millor el dataset. Per tal de realitzar això, per defecte realitza un oversampling d'un 80 % i la resta acaba d'equilibrar-ho amb un undersampling. Això pot ser particularment útil en datasets petits, on cada observació és valiosa. En el cas del projecte moltes vegades es treballaran amb dataset molts petits, per exemple, a l'entrenar els models durant el `Cross-Validation` on les dades estaran molt reduïdes.

2.9.2 Consideracions del Balanceig de Dades

1. **Undersampling:** Realitzar undersampling solament pot no ser pràctic ja que reduiria el nombre d'observacions a aproximadament 20 per classe, resultant en una quantitat de dades insuficient per a l'entrenament dels models.
2. **Oversampling:** L'ús d'oversampling amb SMOTE podria comportar una sobreestimació del rendiment dels models, ja que aproximadament 1/4 de les dades serien sintètiques, afectant la seva qualitat i potencialment la validesa dels resultats.
3. **Combinació d'Oversampling i Undersampling:** Aquesta tècnica podria oferir un equilibri millor, augmentant la mida de les classes minoritàries sense inventar-se massa dades i sense reduir excessivament la quantitat de dades disponibles.
4. **Preferència dels Models per Dades No Balancejades:** És possible que els models prefereixin treballar amb dades no balancejades degut als problemes associats amb les dades sintètiques i la pèrdua d'informació a causa de l'undersampling.

2.9.3 Conclusió

El balanceig de dades és una decisió que depèn del context i de la naturalesa dels models utilitzats. En aquest cas, la combinació d'oversampling i undersampling sembla ser l'opció més viable, però es recomana avaluar l'impacte de cada tècnica en el rendiment del model abans de prendre una decisió final.

3 Preparació de variables

En aquesta secció es comentaran les diferents preparacions que s'han fet a les variables del dataset per tal de poder ajudar i millorar la qualitat de predicció dels models.

S'estudiarà la normalització de variables, equilibrant així el seu pes en alguns algoritmes, les correlacions entre elles juntament amb un estudi bivariat amb la variable objectiu. Tot això per poder trobar quines variables són realment útils i estan relacionades amb la variable *Status* i quines no i, per tant, convé considerar-les de forma diferent ja que podrien incloure soroll en els models.

Per acabar d'entendre i centrar-nos en les variables numèriques s'ha realitzat un estudi de la dimensionalitat mitjançant un **Anàlisi de Components Principals** per tal de poder veure si es pot reduïr el nombre de components numèriques del dataset o si realment totes les variables són necessàries i aporten informació.

Tot això és crucial degut al nombre reduit de mostres amb les que comptem per entrenar models.

3.1 Normalització de variables

La normalització de les variables és crucial en molts algoritmes d'aprenentatge automàtic, ja que assegura que totes les característiques contribueixen equitativament al procés d'aprenentatge, evitant que característiques amb rangs més grans dominin el model.

3.1.1 Elecció de StandardScaler

L'ús de 'StandardScaler', que normalitza les característiques mitjançant la seva mitjana i desviació estàndard, es prefereix sobre 'MinMaxScaler' pels següents motius:

1. **Preservació de la Forma de la Distribució:** 'StandardScaler' manté la forma de la distribució original de la variable, a diferència de 'MinMaxScaler', que pot alterar aquesta forma. Això és particularment important per models com SVM i EBM, que són sensibles a la distribució de les dades.

2. **Robustesa a Valors Extrems:** 'StandardScaler' és menys sensible a valors extrems en comparació amb 'MinMaxScaler'. Això resulta crucial en algoritmes com KNN i SVM, on valors extrems poden tenir un impacte significatiu en la frontera de decisió.

3. **Escala Centrada al Voltant de Zero:** Molts algoritmes, com el SVM i els models basats en arbres (Decision Tree, RandomForest, XGBoost), es beneficien d'una escala centrada al voltant de zero, ja que facilita la convergència i la interpretació dels models.

4. **Compatibilitat amb Diferents Models:** 'StandardScaler' ofereix una major compatibilitat amb una varietat de models utilitzats. Per exemple, KNN i SVM requereixen que totes les característiques tinguin la mateixa escala, mentre que models basats en arbres com Decision Tree i RandomForest són menys sensibles a l'escala però poden beneficiar-se d'una normalització consistent en conjunts de dades amb múltiples característiques de diferents escales.

3.1.2 Distribució de les variables numèriques després de la normalització

Com es pot veure en la figura 14 i al comparar-ho amb les distribucions observades en la secció 2.2.1 les gràfies són pràcticament idèntiques degut a que simplement hem centrat la mitjana a 0 i aplicat una desviació estàndar d'1.

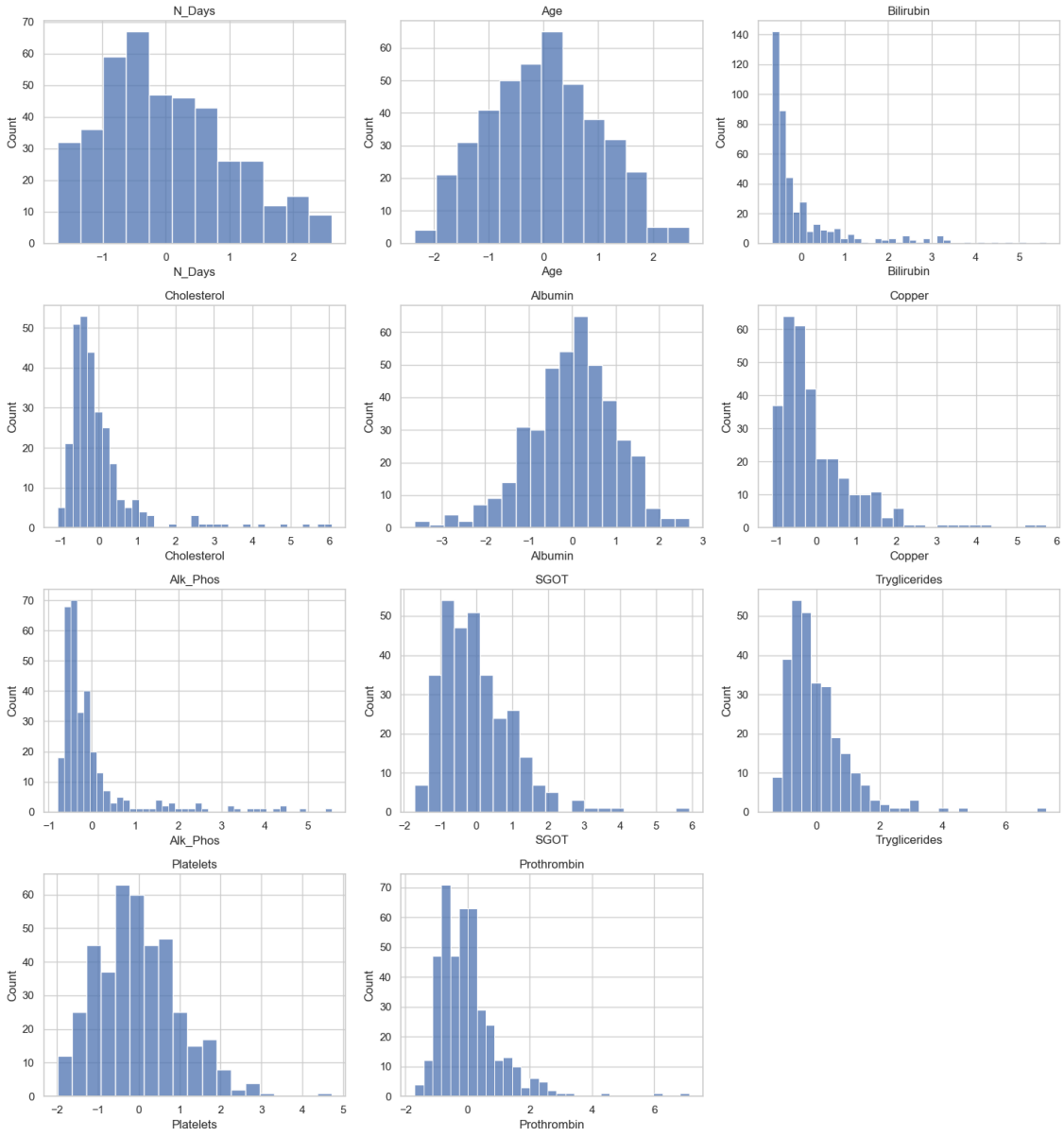

Figure 14: Imatge de les distribucions de les variables numèriques després d'aplicar `StandardScaler`

En conclusió, la normalització de les variables mitjançant 'StandardScaler' és una estratègia eficient per a la preparació de dades en un rang ampli de models d'aprenentatge automàtic, proporcionant un equilibri entre la preservació de la forma de la distribució, la robustesa, i la millora del rendiment del model.

No obstant això la decisió de si normalitzar o no es realitzarà específicament a l'hora d'entrenar un model en específic comparant els seus resultats. Això és degut a que encara que generalment normalitzar ajudar a la majoria de models, potser a l'entrenar un arbre de decisió li vam millor si les dades no es troben escalades.

3.2 Anàlisi de correlacions

S'ha realitzat un estudi estadístic per identificar les correlacions significatives entre diverses variables clíniques utilitzant el coeficient de correlació de Pearson.

La detecció de correlacions entre variables pot ser crucial en el machine learning on correlacions altes poden indicar redundància entre variables, que podria resultar en multicolinealitat en models com regressió lineal. Això pot afectar la capacitat del model per a generalitzar correctament a noves dades.
A més, mentre que algoritmes com arbres de decisió i ensembles basats en arbres (com Random Forest i XGBoost) són menys sensibles a la multicolinealitat, models com SVM o KNN poden beneficiar-se de la reducció de dimensions que elimina variables correlacionades.

3.2.1 Mètode

Mitjançant l'ús de l'funció `pearsonr` de SciPy i iterant sobre totes les combinacions possibles de variables numèriques, s'han calculat les correlacions i els seus respectius valors p. S'han considerat significatives aquelles correlacions amb un valor p inferior a 0.05. A més a més per tal de poder visualitzar gràficament els resultats s'ha creat una matriu de correlació com es pot veure en la figura 15.

3.2.2 Resultats

La taula següent mostra les parelles de variables més significativament correlacionades:

Variable 1	Variable 2	Correlació	Valor p
Bilirubin	Copper	0.486429	$3.256319e^{-26}$
Bilirubin	Tryglicerides	0.464358	$9.597327e^{-24}$
Bilirubin	SGOT	0.463859	$1.086121e^{-23}$
Bilirubin	Cholesterol	0.444893	$1.037343e^{-21}$
N_Days	Albumin	0.430829	$2.547812e^{-20}$

Table 11: Parelles de Variables Més Correlacionades

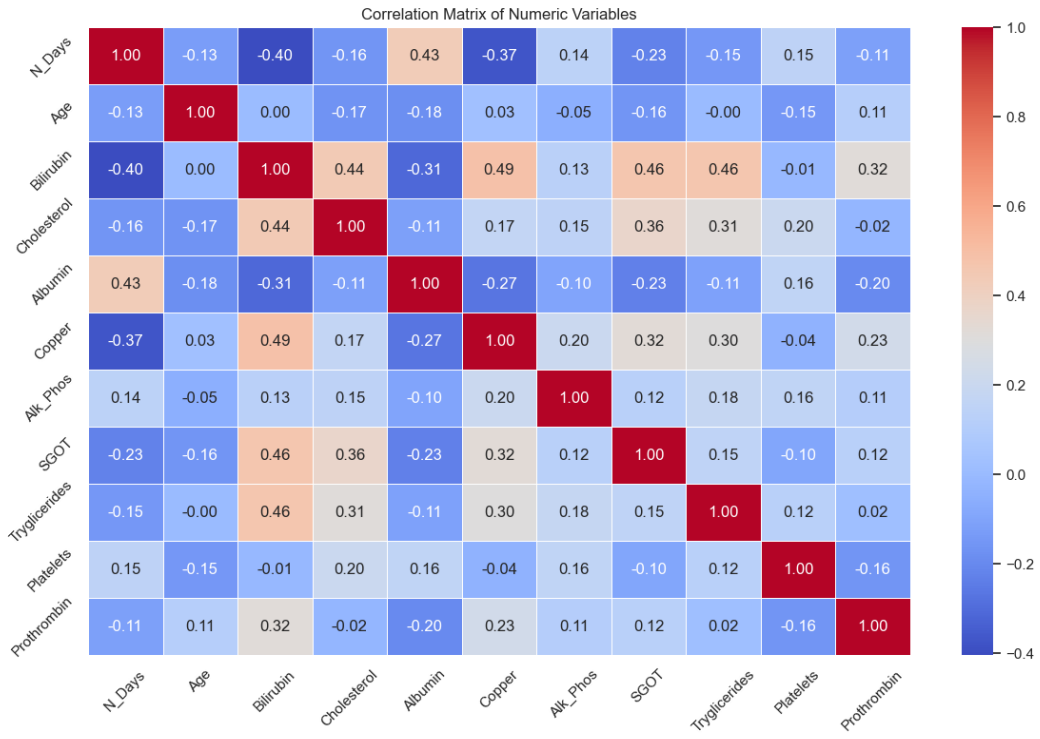

Figure 15: Imatge de la `correlation matrix` de les variables numèriques

En aquesta matriu de correlació 15 es poden observar que les correlacions més fortes suggereixen relacions significatives entre les durades dels períodes d'estudi (`N_Days`) i els nivells d'Albumina, així com entre els nivells de Bilirubina i Copper, i Bilirubina i SGOT. Es mostra una correlació negativa entre els dies i els nivells de Bilirubina i Copper, suggerint que a majors períodes d'estudi, es poden observar nivells més baixos d'aquestes substàncies. També destaca una força linealitat entre la bilirubina i les varaiables Tryglicerides i Cholesterol.

Les implicacions clíniques d'aquestes correlacions requereixen un millor ànalisi, però els resultats podrien indicar efectes patològics compartits o respostes corporals a la malaltia que aquests biomarcadors podrien estar indicant.

3.3 Anàlisi bivariat entre les variables i l'objectiu

L'anàlisi bivariant és una eina crucial en la fase d'exploració de dades d'un projecte de machine learning, especialment quan s'investiga la relació entre característiques independents i la variable objectiu.

3.3.1 Anàlisi bivariat entre les variables numèriques i l'objectiu

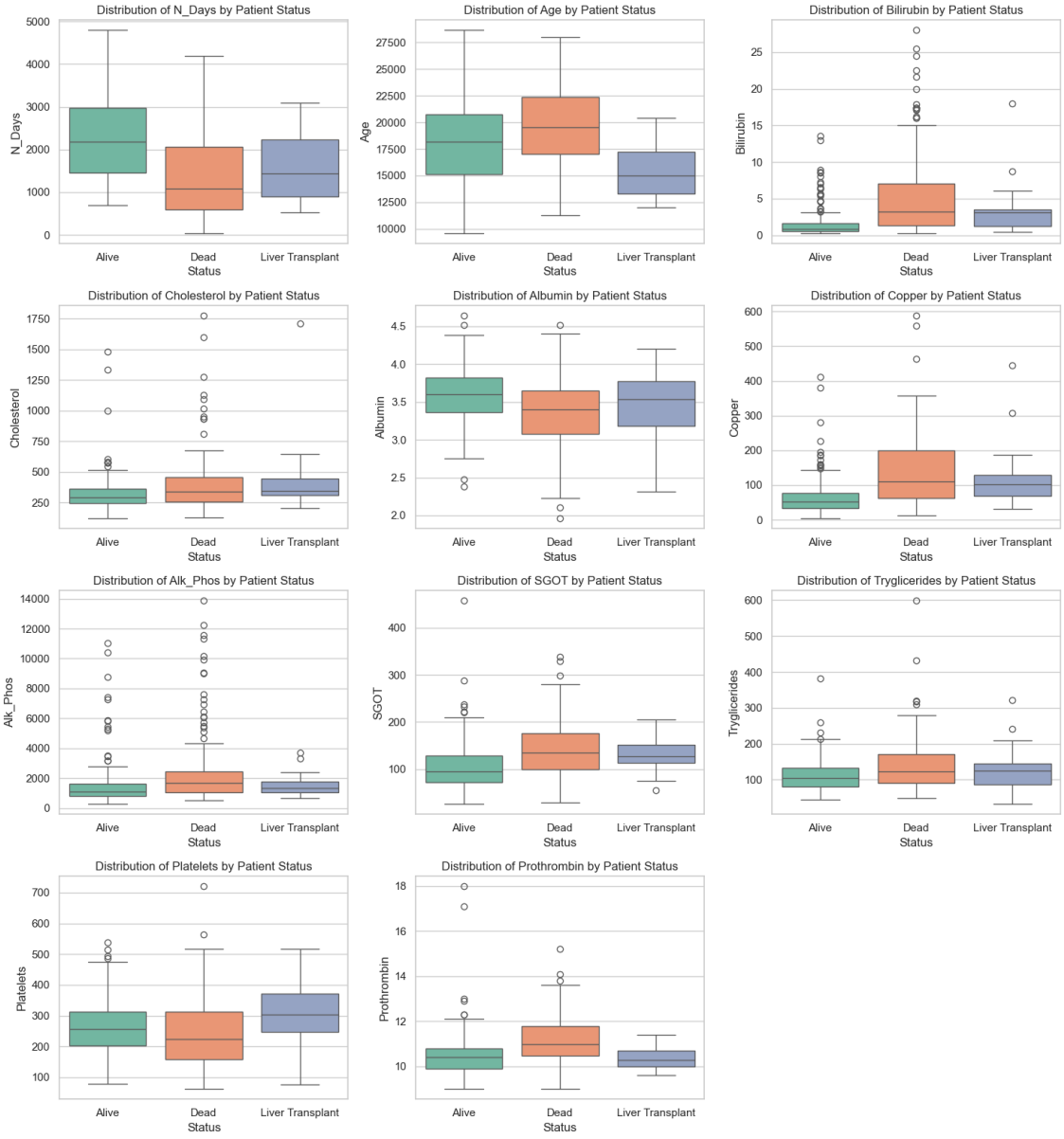

Figure 16: Imatge dels boxplots entre les variables numèriques i la variable objectiu `Status`

Tal com es pot observar en la figura 16, els *boxplots* mostren la distribució de diverses variables biomèdiques respecte a la variable objectiu `Status`, que categoritza l'estat dels pacients com **Alive**, **Dead** o **Liver Transplant**.

Dels boxplots podem treure la següent informació destacable:

- La distribució de `N_Days` també ens aporta informació degut a que la majora de persones que van morir no van estar gaires dies en l'experiment comparat amb els que van tenir un transplantament o que van sobreviure.
- Els nivells de `Bilirubin` mostren diferències notables entre els grups 'Alive', 'Dead' i 'Liver Transplant', suggerint que podrien ser predictius de l'estat del pacient. En concret els pacients que han sobreviscut tenen una concentració de bilirubina molt més propera al 0 que els que no. Tot i que podem trobar outliers també en les persones que han sobreviscut, aquests es considerarien normals en el cas que haguessin mort; indicant així una forta relació entre la bilirubina i la variable objectiu.
- L'ample rang interquartil del `Copper` i els outliers en el grup 'Dead' podrien indicar una associació amb mortalitat més elevada, el que pot ser rellevant per a la predicció de risc.
- El temps de `Prothrombin` també juga un paper clau degut a que la mitjana és lleugerament més eleveda en el cas de que el pacient no sobrevisqués.

3.3.2 Anàlisi bivariat entre les variables categòriques i l'objectiu

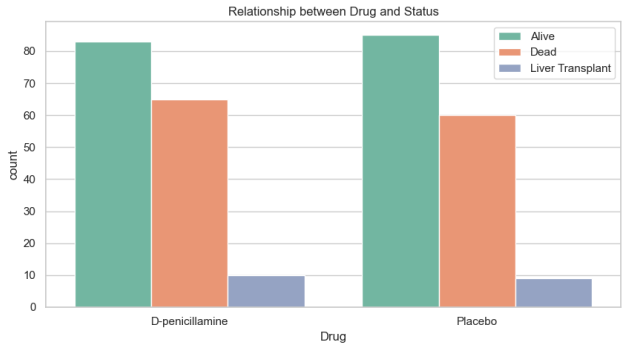

Figure 17: Imatge de la relació entre la variable categòrica `Drug` i l'objectiu `Status`

Com es pot veure en la figura 17, els gràfics de barres que comparen l'estat dels pacients amb el tipus de tractament rebut mostren un patró pràcticament idèntic entre els grups 'D-penicillamine' i 'Placebo'.

Aquest fet provoca les següents implicacions pel que fa a l'estadística i aprenentatge automàtic.

1. La similitud en els gràfics suggereix que no hi ha una diferència significativa en els resultats dels pacients entre els dos tractaments.
2. Aquest resultat podria implicar que el tractament no és un bon predictor de l'estat del pacient, el que redueix la seva utilitat com a característica en un model predictiu.
3. Els models d'aprenentatge automàtic podrien no beneficiar-se significativament de la inclusió d'aquesta variable si no aporta una capacitat discriminatòria addicional.

Aquesta informació ens és molt útil degut a que probablement no aporti gaire informació els models o inclús soroll i, per tant, eliminar o no aquesta variable és un punt interessant de discussió.

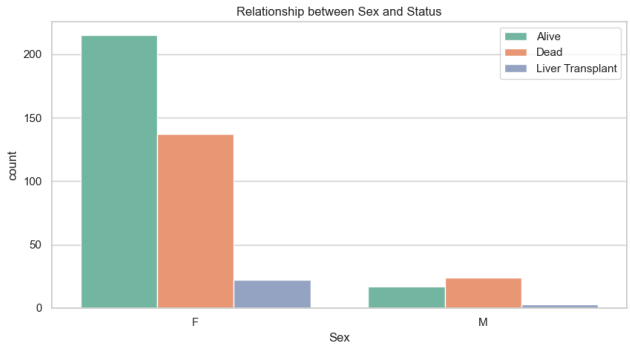

Figure 18: Imatge de la relació entre la variable categòrica `Sex` i l'objectiu `Status`

Pel que fa a la variable `Sex`, en els barplots de la figura 18 es pot observar que com ja s'ha vist anteriorment hi ha moltes més dones que homes i, curiosament, la distribució és força diferent.

En el cas de les dones, la majoria d'elles han sobreviscut de la malaltia però en el cas dels nois la majoria han mort. Tot i el gran desbalanceig que pateix aquesta variable, pot ser interessant utilitzar-la en el models degut a aquesta capacitat discriminatòria que hi trobem.

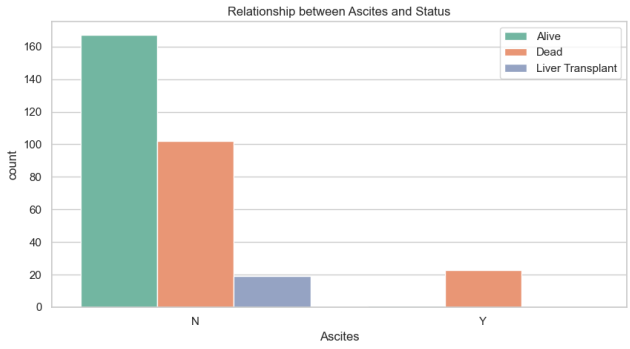

Figure 19: Imatge de la relació entre la variable categòrica `Ascites` i l'objectiu `Status`

En la variable `Ascites` observem un efecte molt curiós. Tal i com es pot observar en la imatge 19 tots els pacients que ho patien, tot i que no són gaires, han mort. Dels que no ho tenien la proporció ja varia una mica degut a que més de 160 han sobreviscut, uns 100 han mort i la resta han tingut un transplant.

Com s'ha comentat també en la variable `Sex`, aquestes diferències de distribució poden resultar

molt útils per a algun models, per exemple arbres de decisió, per a predir la variable objectiu.

Figure 20: Imatge de la relació entre la variable categòrica `Hepatomegaly` i l'objectiu `Status`

En la variable `Hepatomegaly` també hi trobem diferències considerables, tal i com es pot observar en la imatge 20, on els pacients que ho pateixen tenen moltes més probabilitats de morir que els que no.

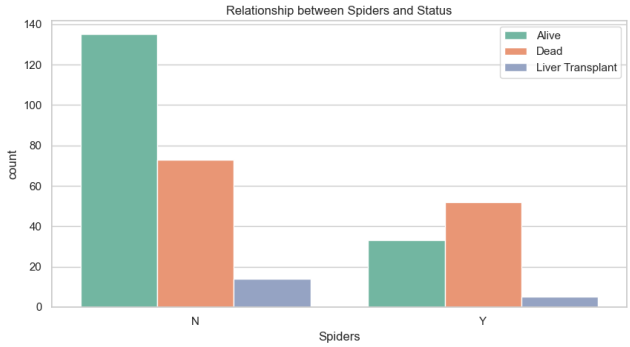

Figure 21: Imatge de la relació entre la variable categòrica `Spiders` i l'objectiu `Status`

Pel que fa a la variable `Spiders` (figura 21) els malalts que la pateixen també tenen unes probabilitats més elevades de no sobreviure a la *Cirrosis* comparats amb els que no la pateixen.

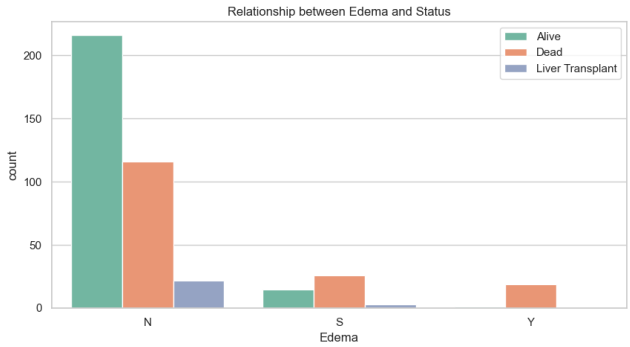

Figure 22: Imatge de la relació entre la variable categòrica `Edema` i l'objectiu `Status`

Si analitzem els barplots de la variable *Edema* (figura 22) podem extreure que els pacients que no tenen Edema tenen moltes més probabilitats de sobreviure que els que la tenen. A més a més també tenen una molt major probabilitat de rebre un transplantament de fetge.

Concretament els que tenen un edema resolt o sense diürètics observem que encara que tinguin més possibilitats de morir que de sobreviure, la proporció vindria a ser al voltant d'un 60% mentre que els que tenen un edema i tractament diürètic estan morts tots. Aquesta relació molt porbablement és deguda a la gravetat de la salut del pacient no tan sols pel que fa a la cirrosis sinó també en general. Aquesta informació pot resultar particularment útil per als models.

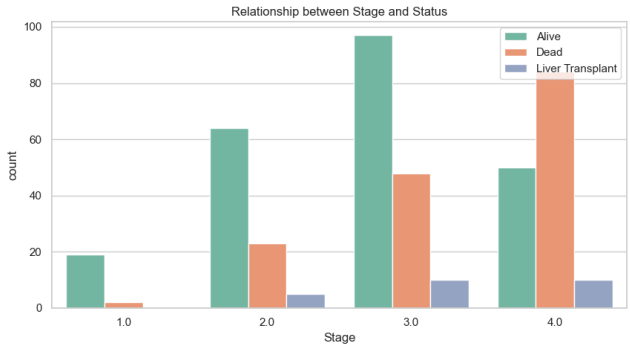

Figure 23: Imatge de la relació entre la variable categòrica `Stage` i l'objectiu `Status`

Finalment en la figura 23 podem treure conclusions molt importants.

- Etapa 1: Hi ha pocs pacients en aquesta categoria, el que pot indicar una etapa inicial de la malaltia. Els pacients aquí estan majoritàriament vius, indicant que aquesta etapa pot estar associada amb millors resultats de salut.
- Etapa 2: Aquesta etapa mostra una alta incidència de mortalitat. Això suggeriria que és un punt crític en la progressió de la malaltia, on els models predictius podrien centrar-se per a identificar pacients en risc.

- Etapa 3: La majoria dels pacients en aquesta categoria estan vius, però també hi ha un nombre significatiu de trasplantaments de fetge, suggerint que l'etapa pot estar associada amb una malaltia moderadament avançada.
- Etapa 4: Aquest última etapa mostra el major nombre de trasplantaments de fetge, el que indica una condició avançada que requereix una intervenció mèdica significativa. A més a més és la única etapa en la qual el nombre de morts és superior al de vius fet que podrien considerar els models alhora de ser entrenats.

3.3.3 Conclusions

Els resultats de l'anàlisi bivariant entre les variables independents i la variable objectiu `Status` ofereixen una visió diversificada que és fonamental per al desenvolupament de models predictius en machine learning. Les conclusions que podem extreure d'aquesta anàlisi són les següents:

1. Les variables biomèdiques com `Bilirubin`, `Copper` i `Prothrombin` mostren diferències significatives entre els pacients segons el seu estat (viu, mort, trasplantament de fetge) i poden actuar com a forts predictors dins d'un model predictiu.
2. La variable `Drug`, que no mostra una diferència significativa entre els grups de tractament, pot no ser tan útil per als models predictius, tot i que la seva inclusió o exclusió ha de ser validada a través de tècniques de selecció de característiques i modelització.
3. Altres variables categòriques, com `Sex`, `Ascites`, `Hepatomegaly`, `Spiders` i `Edema`, mostren distribucions que suggereixen una capacitat discriminatòria i poden ser especialment valuoses per a models com arbres de decisió que poden capturar la seva influència en el `Status`.
4. La variable `Stage` demostra ser crucial, amb diferents estadis de la malaltia associats amb resultats clínics específics. Aquesta variable pot ajudar a identificar pacients amb un risc més elevat de mort o necessitat de trasplantament de fetge.

La integració d'aquestes variables i la comprensió de les seves relacions amb la variable objectiu és clau per al desenvolupament de models predictius eficaços. Cal realitzar una validació creuada i ajustar els models per a les particularitats del conjunt de dades per assegurar que els models són robustos, precisos i capaços de generalitzar a nous casos.

3.4 Eliminació de variables redundants o sorolloses

Durant el pre-processament de dades, és essencial identificar i eliminar les variables que no aporten informació útil al model. Aquestes poden ser variables redundants, que no afegeixen cap nou coneixement degut a la seva alta correlació amb altres característiques, o variables sorolloses, que poden introduir variabilitat innecessària als models predictius i, per tant, reduir la seva capacitat per generalitzar a noves dades.

1. **Anàlisi de Correlació**: Com a primer pas, cal realitzar un anàlisi de correlació per a totes les parelles de variables contínues. Les variables amb una alta correlació ($r > 0.9$) són candidates per a la seva eliminació però com ja s'ha observat en el nostre dataset el màxim que trobem és 0.4 i per tant no tenim evidència suficient per eliminar cap variable numèrica.

2. **Explainable Boostin Machine**: Com es pot observar en el *Bouns 1*, hi ha variables que aporten poca informació al model així que podrien considerar-se candidates per a la seva eliminació.

3. **Validació Creuada**: Abans de prendre una decisió final sobre la eliminació d'una variable, realitzem una validació creuada per assegurar que la seva eliminació millora o manté la precisió del model. Aquest pas és crucial per evitar l'eliminació de variables que poden semblar no informatives de forma aïllada, però que en combinació amb altres poden aportar informació valuosa.

La decisió final sobre quines variables eliminar es pren després d'una consideració detallada de l'impacte que tindran en el rendiment del model, sempre amb l'objectiu de simplificar el model sense sacrificar la seva capacitat predictiva. En el cas de la nostra base de dades, degut a que disposem de molt poques observacions, s'ha decidit no eliminar cap variable per molt poca variància o informació que pugui aportar.

3.5 Estudi de la dimensionalitat

En aquest estudi s'ha aplicat una tècnica d'Anàlisi de Components Principals (`PCA`) per reduir la dimensionalitat del conjunt de dades de pacients amb cirrosi i identificar les característiques més significatives que poden influir en la predicció de la supervivència dels pacients, la mort o la necessitat d'un trasplantament de fetge.

S'ha realitzat la imputació de les dades per tractar els valors perduts. Posteriorment, es van estandarditzar les dades numèriques per tal de tenir una mitjana de zero i una desviació estàndard d'1. Això és essencial per a PCA, ja que el mètode és sensible a les escales de les variables. Després, s'ha aplicat PCA als dades estandarditzades per transformar-les en un conjunt de valors de components principals.

La Figura 24 mostra la variància explicada acumulativa. Es pot observar que aproximadament 7 components principals són necessaris per explicar el 80% de la variància. Aquest resultat ens indica que amb tan sols 7 variables podem explicar gairebé tota la variància pel que fa a les variables numèriques. Cal considerar, però, que aquest 20% restant pot ser molt important i més en el nostre dataset dels quals qualsevol mínima informació és necessària degut a la poca quantitat d'observacions.

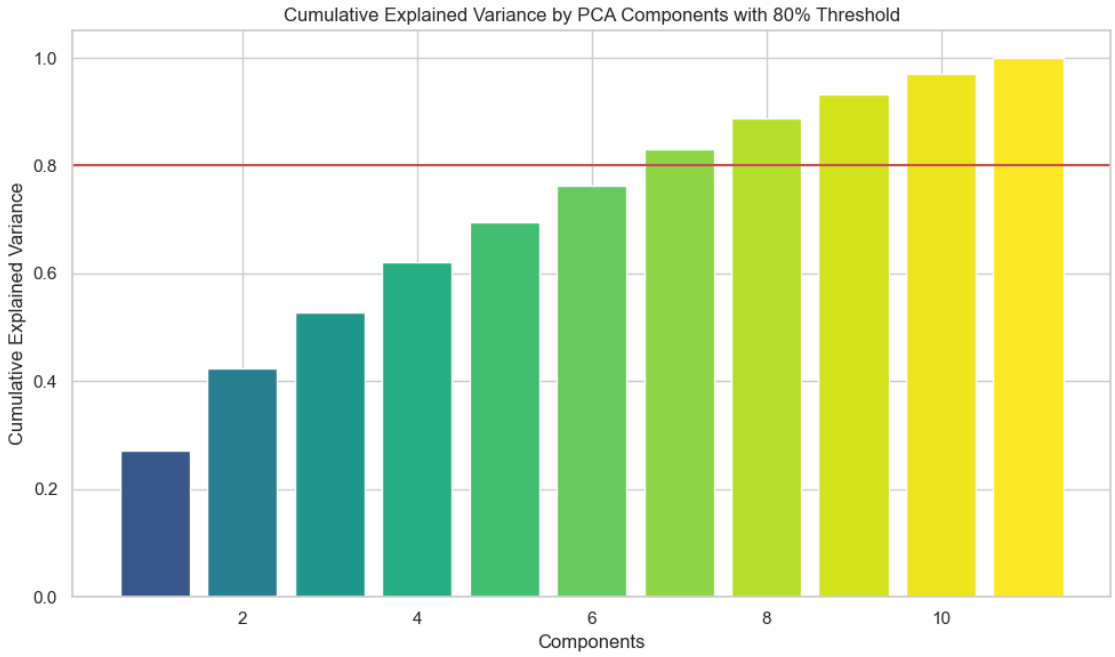

Figure 24: Imatge de l'acumulació de la variància en els components

En la Figura 25, cada vector representa una variable numèrica. La direcció i longitud del vector indiquen com aquesta variable contribueix a cada component principal.

Com que les variables es troben projectades sobre els dos primer components, i com es pot veure en la figura 24, estariem observant un 40% de tota la variància així que les conclusions que es poden treure no són realment del tot fiable però ens poden indicar i ajudar a entendre com es relacionen les variables numèriques.

El primer component principal, l'eix **X**, es troba controlat positivament per la variable `Bilirubin`, `Copper` i `SGOT` i negativament per les variables `N_Days` i `Albumin`. Aquesta informació no és nova degut a que ja havíem observat aquestes correlacions quan s'ha realitzat l'estudi de correlacions anteriorment.

Això ens indica que les persones que tenen més `Bilirubin` alhora que `Copper` i `SGOT` de normal estan relacionats a tenir menys dies en l'experiment, molt probablement degut a que no acabin sobrevivint.

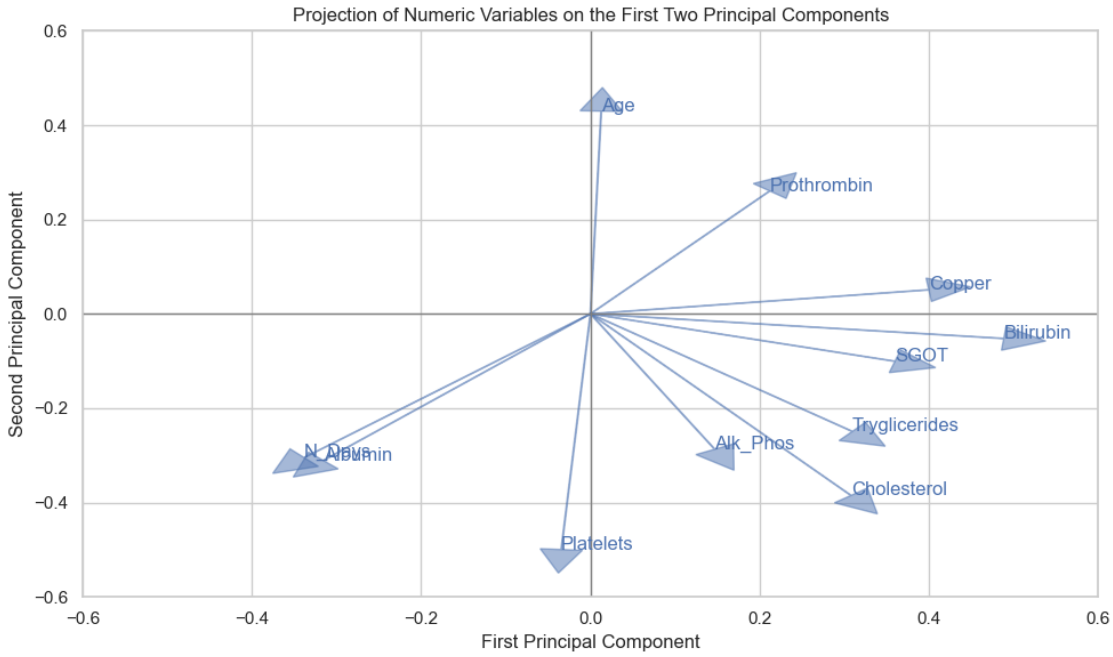

Figure 25: Imatge de les projeccions de les variables numèriques sobre els dos primers components

Pel que fa al segon component principal, l'eix $\mathbf{Y}$, es troba controlat positivament per la variable `Age` i negativament per la variable `Platelets` i `Cholesterol`. Basant-nos en aquesta visualització, podem inferir que a mesura que augmenta l'edat, els nivells de plaquetes i colesterol tendeixen a disminuir. Això pot ser degut a diversos factors biològics i fisiològics associats amb l'envelliment com ara a que la funció hepàtica pot disminuir en persones grans i també a que la mèdula òssia pot tornar-se menys eficient en la producció de cèl·lules sanguínies incloses les plaquetes.

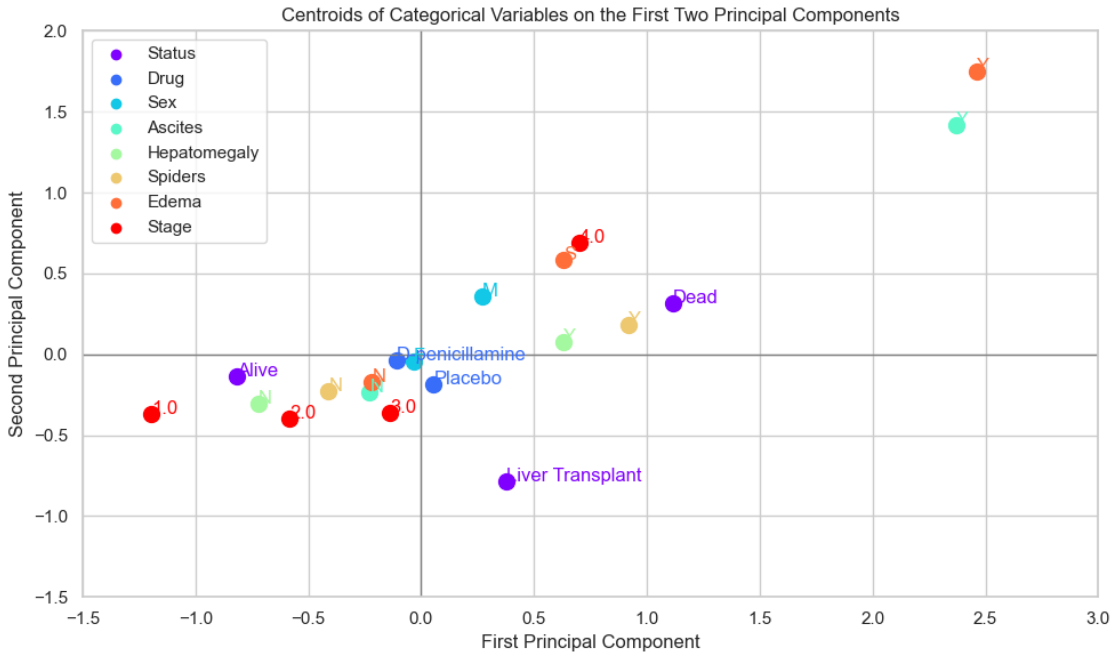

Figure 26: Imatge de la projecció dels centroides de les varibles categòriques sobre els primers 2 components

La figura 26 dels centroides en l'espai dels dos primers components principals mostra la mitjana dels valors de les observacions que pertanyen a cada categoria de les variables categòriques. Els centroides són calculats com la mitjana en l'espai reduït de PCA, que són les dues dimensions principals que expliquen la major variància en les dades. En aquest cas, cada punt representa la mitjana dels valors de PCA per a totes les observacions que comparteixen una categoria específica, agrupades per colors. Aquestes projeccions ens ajuden a entendre com les categories diferents estan distribuïdes en l'espai reduït.

De totes les variables categòriques destaquen les següents:

- `Status`: La variable objectiu es troba molt ben representada degut a que els centroides de les seves classes estan molt distanciats. La classe *Alive* es troba en la banda esquerra de l'eix **X** indicant valors baixos en bilirubina, coure però valors més alts pel que fa al nombre de dies que es troben en l'experiment i l'albumina. La classe *Liver Transplant* es troba amb valors més elevats de bilirrubina però en aquest cas el trobem en el 4 quartil indicant un nivell alt de plaquetes i una edat més jove; probablement sigui degut a que les persones a les quals feien un transplantament no podien ser massa grans i debien tindre algun problema extra en la sang. Per últim la categoria *Dead* es pot observar en la banda dreta i a dalt indicant que són persones amb uns valors clínics més elevats, és a dir, que es troben més malament i d'edat avançada.
- `Stage`: Com era d'esperar aquesta variable està força relacionada amb l'objectiu. En el cas

que una persona estigui en l'etapa 1 de la malaltia, aquesta estarà en la banda esquerra del gràfic i a mesura que l'etapa de la malaltia va avançant, el pacient estaria més cap a la dreta. En la última etapa, la quarta, els malalts es troben també en la part superior indicant que podrien ser persones d'edat més reduïda.

- `Scites`: Com ja s'havia vist en l'estudi bivariat de les variables amb l'objectiu, aquesta variable en el cas que un pacient la pateixi té un 100% de mortalitat. Això es troba representat en aquests dos components principals degut a que en el cas que un pacient no ho pateixi es troba centrat a l'eix de coordenades però en el cas que sí és de les categories més allunyades cap a la dreta i a dalt.
- `Edema`: Passa una cosa similar en el cas en que un pacient tingui o no un edema i també depen del tipus. En el cas que no en tingui es troba centrat en l'eix de coordenades. En el cas que tingui un edema resolt o sense diürètics es troba més cap a la dreta i a dalt. Finalment en el cas que tingui un edema com també s'ha observat en l'estudi bivariat la mortalitat està assegurada i, per tant, el centroide està ubicat en la posició més allunyada cap a la dreta i a dalt.

Finalment, tal com ja s'havia comentat, en el nostre dataset no ens interessa perdre el 20% de la variància encara que reduïm considerablement el nombre de dimensions. A més a més perdriem molta interpretabilitat alhora que degut a la poca quantitat d'observacions que tenim ens podem permetre que el cost d'entrenar models sigui més elevat.

4 Models

En aquest apartat es discutiran els models entrenats per tal de poder predir la variable objectiu. En ells es realitzarà un estudi de les mètriques de rendiment, la selecció d'hiperparàmetres i finalment l'anàlisi dels resultats.

4.1 Funcions utilitzades

Importacions

- `from sklearn.model_selection import train_test_split, StratifiedKFold`: Aquestes funcions s'utilitzen per dividir el conjunt de dades en conjunts d'entrenament i prova (`train_test_split`) i per a realitzar validació creuada estratificada (`StratifiedKFold`).
- `from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score`: Aquestes funcions calculen diferents mètriques d'avaluació com la precisió, la puntuació F1, la precisió i la recuperació.
- `from sklearn.preprocessing import StandardScaler, OneHotEncoder, OrdinalEncoder, LabelEncoder`: Aquestes funcions s'utilitzen per preprocessar dades, incloent l'escalat de característiques numèriques (`StandardScaler`) i la codificació de característiques categòriques (`OneHotEncoder`, `OrdinalEncoder`, `LabelEncoder`).
- `from xgboost import XGBClassifier`: Importa el classificador XGBoost, un model potent i eficient per a tasques de classificació.

Funció `cross_validate_model`

Aquesta funció realitza la validació creuada d'un model donat, utilitzant diferents tècniques de preprocessament i equilibrat de conjunts de dades. Utilitza `StratifiedKFold` per dividir les dades, i aplica codificació, normalització, i mètodes de equilibri si es requereixen. Finalment, avalua el model utilitzant mètriques com la precisió, puntuació F1, precisió i recuperació.

Funció `train_model`

Aquesta funció entrena un model utilitzant un conjunt de dades donat. Permet l'ajust de hiperparàmetres, aplicació de preprocessament, i equilibrat de les dades. També realitza una validació creuada i avalua el model en el conjunt de prova, proporcionant mètriques de rendiment.

Funció `plot_confusion_matrix`

Aquesta funció visualitza una matriu de confusió, que és útil per entendre el rendiment del model en termes de les diferents classes.

Funció `plot_multiclass_roc_curve`

Aquesta funció dibuixa la corba ROC per a cada classe en problemes de classificació multiclasse, proporcionant una visualització de la capacitat del model per distingir entre les classes.

Funció `find_best_dataset_combination`

Aquesta funció explora diferents combinacions de preprocessament i ajustaments en un conjunt de dades per trobar la millor configuració per a diferents models. Utilitza validació creuada per a avaluar cada combinació.

4.1.1 Observacions del dataset

S'ha utilitzat la funció `find_best_dataset_combination` per explorar diverses combinacions de modificacions, incloent l'eliminació de files, la eliminació d'outliers, la codificació ordinal, l'oversampling, i la normalització.

Durant l'anàlisi, s'ha observat que l'eliminació d'una quantitat substancial de dades, com ara les últimes 106 files d'un conjunt de dades de 400, pot reduir significativament la varietat de casos disponibles per a l'aprenentatge del model. Això podria comprometre la capacitat de generalització del model, especialment en casos clínics on els valors alts de bilirrubina són freqüents i rellevants. A més, la combinació de l'eliminació massiva de dades amb l'aplicació d'oversampling pot resultar en un model entrenat predominantment amb dades sintètiques, el que pot afectar negativament la seva capacitat per interpretar dades reals.

A més a més en tots els models en els que s'utilitzava la funció els millors resultats sempre s'obtenien d'eliminar tantes dades com era possible.

Conseqüentment, s'ha decidit utilitzar els resultats de la funció principalment per determinar la necessitat de normalitzar les dades, equilibrar les classes, i seleccionar el tipus de codificació més adequat. Aquest enfocament assegura que el preprocessament de les dades no comprometi la diversitat i representativitat del conjunt de dades original, mantenint així la integritat de l'anàlisi i la validesa dels models en l'àmbit de la salut.

4.2 KNearest Neighbors Classifier (KNN)

L'aplicació del KNN en aquest estudi té com a objectiu proporcionar un model que pugui classificar els pacients segons la seva probabilitat de supervivència, mort o necessitat d'un trasplantament de fetge, basant-se en les seves característiques clíniques.

4.2.1 Mètriques de Rendiment

Les mètriques de rendiment seleccionades són crucials per a la validació de la robustesa i fiabilitat del model:

- **Precisió (Precision)**: Mesura la proporció de prediccions positives correctes entre totes les prediccions positives. És important quan el cost de falsos positius és alt.
- **Accuradesa (Accuracy)**: Mesura la proporció de prediccions correctes en tot el conjunt de dades. És útil quan les classes estan relativament equilibrades.
- **Recall**: Indica la proporció de casos positius reals que s'han predit correctament.
- **F1-Score**: Combina precisió i recall en una sola mètrica que busca un equilibri entre ambdues.

- **AUC-ROC**: Mesura la capacitat del model de discriminar entre classes. Ideal quan es tracta de classes desequilibrades.

4.2.2 Selecció d'Hiperparàmetres

El model KNN es tria per la seva capacitat de maneig de dades no lineals i la seva flexibilitat en la classificació. A més, és un model interpretable, ja que la classificació es basa en la proximitat als casos més similars. Aquestes característiques són desitjables donada la complexitat i la naturalesa del problema en qüestió.
Els hiperparàmetres considerats per a aquest model són:

- **Número de veïns (k)**: El nombre de veïns més propers a considerar per a la classificació.
- **Pesos**: Determina si tots els veïns contribueixen igualment a la classificació (uniforme) o si els més propers tenen més pes (inversament proporcional a la distància).
- **Mètrica de Distància**: Determina com es calcula la distància entre punts, típicament euclidiana o manhattan.
- **leaf_size**: Afecta l'eficiència de la construcció i consulta de l'estructura de dades interna (per exemple, BallTree o KDTree).

La taula d'hiperparàmetres i valors provats es pot representar com:

Hiperparàmetre	**Valors Provats**
Número de veïns (k)	3, 5, 7, 9
Pesos	Uniforme, Distància
Mètrica de Distància	Euclidiana, Manhattan
leaf_size	20, 30, 40, 50

Table 12: Llista d'hiperparàmetres i valors provats per KNN.

4.2.3 Entrenament i validació

S'ha dividit el conjunt de dades en un 80% per a entrenament i un 20% per a test, juntament amb l'aplicació de la validació creuada per garantir que el model és generalitzable i robust. Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler`
- Encodificar amb `OrdinalEncoder`

Durant la fase d'entrenament, s'han ajustat els hiperparàmetres mitjançant la validació creuada per optimitzar les mètriques de rendiment. Els resultats de cada iteració s'han analitzat per determinar la presència d'overfitting o underfitting i fer els ajustos necessaris.

4.2.4 Anàlisi dels resultats

El resultat final del model KNN, després de seleccionar els millors hiperparàmetres, s'ha registrat en la següent taula 13.

Hiperparàmetre	**Valor**
leaf_size	20
metric	euclidean
n_neighbors	3
weights	uniform

Table 13: Hiperparàmetres optimitzats del model KNN

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts del train en la validació creuada són les següents.

Mètrica	**Valor**
Accuracy	0.8263
F1 Score	0.8165
Precision	0.8287
Recall	0.8263

Table 14: Mètriques mitjanes d'entrenament en la validació creuada

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts de validació dins del cross-validations són les següents.

Mètrica	**Valor**
Accuracy	0.7006
F1 Score	0.6847
Precision	0.6861
Recall	0.7006

Table 15: Mètriques mitjanes de validació en la validació creuada

Com es pot observar, el KNN Classifier pel que fa al cross-validation ha fet overfitting tot i que tampoc ha estat tan exagerat ja que les diferències mitjanes amb les mètriques és de prop del 10%.

Class	Precision	Recall	F1-score	Support
Alive	0.80	0.92	0.86	185
Dead	0.86	0.75	0.80	129
Liver Transplant	0.75	0.30	0.43	20
Accuracy			0.82	334
Macro avg	0.80	0.66	0.70	334
Weighted avg	0.82	0.82	0.81	334

Table 16: Train Classification Report

En la taula 16 es pot observar uns valors lleugerament superiors als que ens trobavem amb la part del train del validation. Això es molt probablement degut a que en aquest train estem entrenant amb el 80% de les dades mentre que en les particions del cross-validation tenim un 80% d'aquest 80% fent que sigui més difícil adaptar-te a les dades sense fer overfitting.

Class	Precision	Recall	F1-score	Support
Alive	0.71	0.87	0.78	47
Dead	0.69	0.56	0.62	32
Liver Transplant	0.00	0.00	0.00	5
Accuracy			0.70	84
Macro avg	0.47	0.48	0.47	84
Weighted avg	0.66	0.70	0.67	84

Table 17: Test Classification Report

Si ens fixem amb el report per part del test en la taula 17 podem veure que els valors no difereixen gaire del train significant que tampoc s'ha fet massa overfitting. D'altra banda cal destacar que no ha predit cap transplantament de fetge correctament cosa que és bastant lògic degut al gran desbalanceig de les dades (encara que si es balancegen els resultats són molt pitjors).

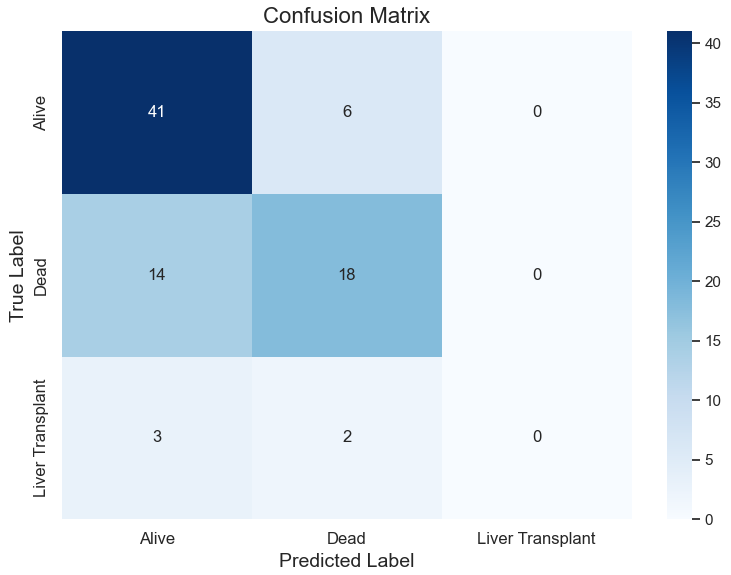

Figure 27: Matriu de confusió del conjunt de test

Pel que fa a la matriu de confusió del test podem veure que a la classe 'Alive', el model ha predit correctament 41 casos, però ha confós 6 casos com a 'Dead'. Per a la classe 'Dead', hi ha 14 casos que s'han predit incorrectament com a 'Alive' i 18 casos han estat correctament classificats. Per a 'Liver Transplant', el model no ha predit correctament cap cas, amb 3 casos classificats com a 'Alive' i 2 com a 'Dead'. Això pot indicar una dificultat del model per reconèixer aquesta classe específica o una falta de dades representatives per a la classe en el conjunt de dades.

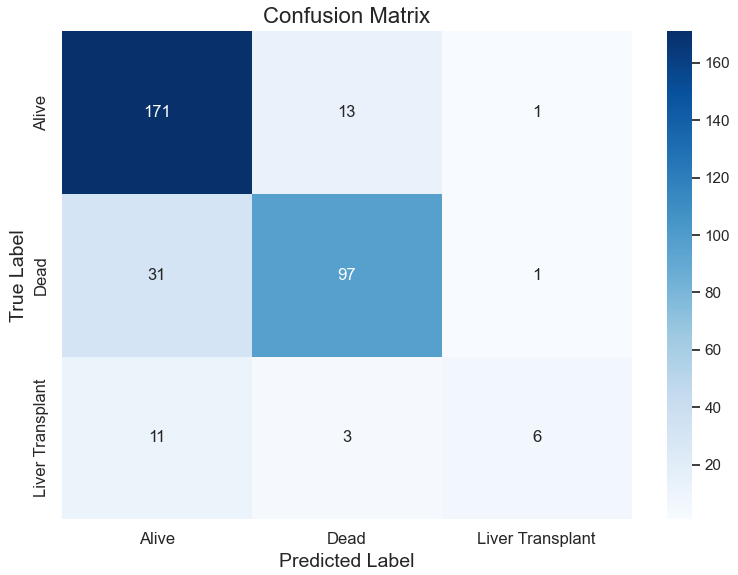

Figure 28: Matriu de confusió del conjunt d'entrenament 'real'

Pel que fa a la matriu de confusió del train podem veure que hi ha una millor classificació general comparada amb el conjunt de test, amb 171 casos d"Alive' i 97 de 'Dead' correctament identificats. No obstant això, encara hi ha confusions, amb 31 casos de 'Dead' incorrectament classificats com a 'Alive' i 11 casos de 'Liver Transplant' com a 'Alive'. Això reafirma la tendència del model a confondre les classes menys representades.

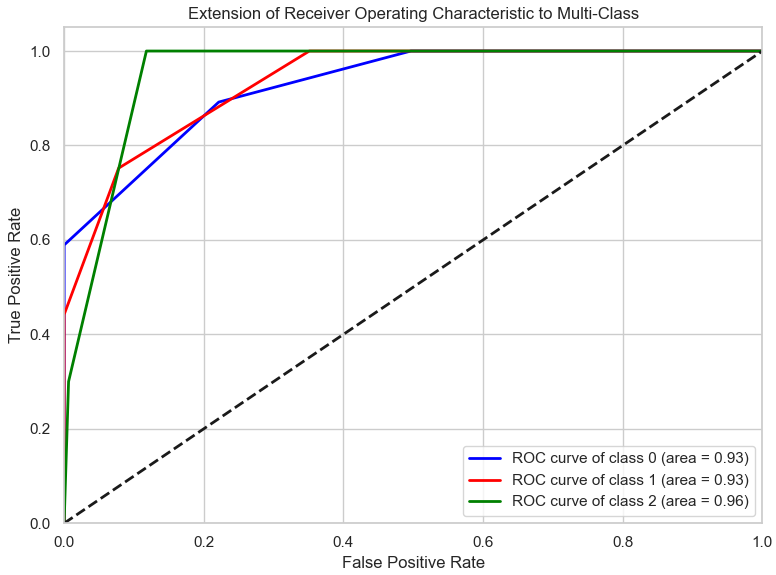

Figure 29: Corba ROC del conjunt d'entrenament

La curva ROC del train mostra que a la classe Alive i a la classe Dead, tenen àrees sota la corba (AUC) de 0.93, indicant una excel·lent capacitat discriminatòria. La classe 'Liver Transplant' té un AUC de 0.96, encara millor, suggerint que el model té una molt bona capacitat per separar aquesta classe específica dels altres.

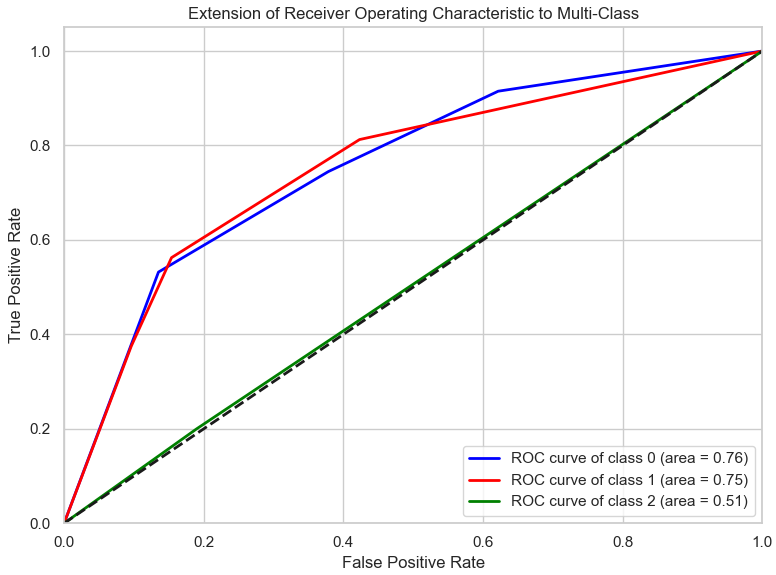

Figure 30: Corba ROC del conjunt de test

Pel que fa a a la curva ROC del test podem observar que és força similar a la del train, però amb una performance lleugerament inferior, com es pot esperar en un conjunt de dades no vist anteriorment. Les àrees sota la corba per a les classes 0 i 1 són 0.76 i 0.75 respectivament, el que indica una bona capacitat predictiva, però no tan alta com en l'entrenament. La classe 2 té un AUC significativament més baix de 0.51, que és gairebé aleatòria, el que podria indicar que el model no ha generalitzat bé per aquesta classe o que hi ha molt poques instàncies per aprendre correctament.

Per concloure l'overfitting és una preocupació aquí, especialment amb respecte a la classe 'Liver Transplant', que sembla ser difícil de predir tant en l'entrenament com en la prova. Aquest model pot no ser adequat per a la predicció d'aquesta classe sense més ajustos o dades addicionals. En un entorn clínic, aquest tipus de model necessitaria ser altament fiable, ja que els falsos negatius o falsos positius poden tenir conseqüències greus. Aquestes mètriques indiquen que, mentre que el model podria ser útil per predir si un pacient estarà 'Alive' o 'Dead', és menys fiable per a la predicció de 'Liver Transplant', i caldria més investigació o ajustos en el model.

4.3 Decision Tree Classifier

Els arbres de decisió són models predictius molt utilitzats que segmenten el conjunt de dades en branques per formar un arbre de decisions. Aquests models són especialment apreciats per la seva interpretabilitat i la seva capacitat per manejar dades no lineals.

4.3.1 Mètriques de Rendiment

Per avaluar l'eficàcia del nostre arbre de decisió, utilitzarem les següents mètriques:

- **Precisió (Precision)**: Important quan el cost de falsos positius és rellevant.

- **Accuradesa (Accuracy)**: Proporció de prediccions correctes sobre el total.
- **Recall**: Especialment crític quan és important capturar tots els casos positius.
- **F1-Score**: Balança entre precisió i recall, útil quan necessitem un compromís entre ambdós.
- **AUC-ROC**: Indica la capacitat del model per distingir entre les classes.

4.3.2 Selecció d'Hiperparàmetres

Els arbres de decisió tenen diversos hiperparàmetres que poden ser ajustats per millorar el rendiment. Els més importants són:

- **max_depth**: La profunditat màxima de l'arbre. Limitar la profunditat pot prevenir sobreajust.
- **min_samples_split**: El nombre mínim de mostres requerides per dividir un node.
- **min_samples_leaf**: El nombre mínim de mostres requerides en un full o node terminal.
- **max_features**: El nombre màxim de característiques a considerar quan es busca la millor divisió.
- **criterion**: La funció per mesurar la qualitat d'una divisió. 'gini' per la impuresa de Gini o 'entropy' per la guany d'informació.

Els valors provats per aquestos hiperparàmetres es poden veure a la Taula 18.

Hiperparàmetre	Valors Provats
max_depth	3, 5, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	sqrt, log2, None
criterion	gini, entropy

Table 18: Llista d'hiperparàmetres i valors provats per l'arbre de decisió.

4.3.3 Entrenament i Validació

Hem dividit el conjunt de dades utilitzant una proporció del 80% per a entrenament i un 20% per a prova, a més d'aplicar la validació creuada per assegurar la generalització del model. Aquest enfocament ens permet ajustar els hiperparàmetres de manera més fiable i identificar si el model està sobreajustant.

Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler`
- Encodificar amb `OrdinalEncoder`

4.3.4 Anàlisi dels resultats

Els resultats del model Decision Tree, després de seleccionar els millors hiperparàmetres, s'han registrat en la següent taula:

Hiperparàmetre	Valor
criterion	entropy
max_depth	10
max_features	None
min_samples_leaf	2
min_samples_split	2

Table 19: Hiperparàmetres optimitzats del model Decision Tree

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts del train en la validació creuada són les següents:

Mètrica	Valor
Accuracy	0.9401
F1 Score	0.9377
Precision	0.9412
Recall	0.9401

Table 20: Mètriques mitjanes d'entrenament en la validació creuada del Decision Tree

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts de validació dins del cross-validation són les següents:

Mètrica	Valor
Accuracy	0.6917
F1 Score	0.6847
Precision	0.6811
Recall	0.6917

Table 21: Mètriques mitjanes de validació en la validació creuada del Decision Tree

Com es pot observar, el Decision Tree Classifier pel que fa al cross-validation ha fet overfitting ja que les diferències mitjanes amb les mètriques és d'aproximadament un 25%.

Class	Precision	Recall	F1-score	Support
Alive	0.96	0.96	0.96	185
Dead	0.92	0.95	0.94	129
Liver Transplant	0.94	0.80	0.86	20
Accuracy			0.95	334
Macro avg	0.94	0.90	0.92	334
Weighted avg	0.95	0.95	0.95	334

Table 22: Train Classification Report del Decision Tree

En la taula 22 es pot observar uns valors molt superiors als que ens trobavem amb la part del train del validation, indicant overfitting del model.

Class	Precision	Recall	F1-score	Support
Alive	0.79	0.70	0.74	47
Dead	0.65	0.75	0.70	32
Liver Transplant	0.00	0.00	0.00	5
Accuracy			0.68	84
Macro avg	0.48	0.48	0.48	84
Weighted avg	0.69	0.68	0.68	84

Table 23: Test Classification Report del Decision Tree

Si ens fixem amb el report per part del test en la taula 23, podem veure que els valors han disminuït en comparació amb l'entrenament, cosa que reafirma l'overfitting observat.

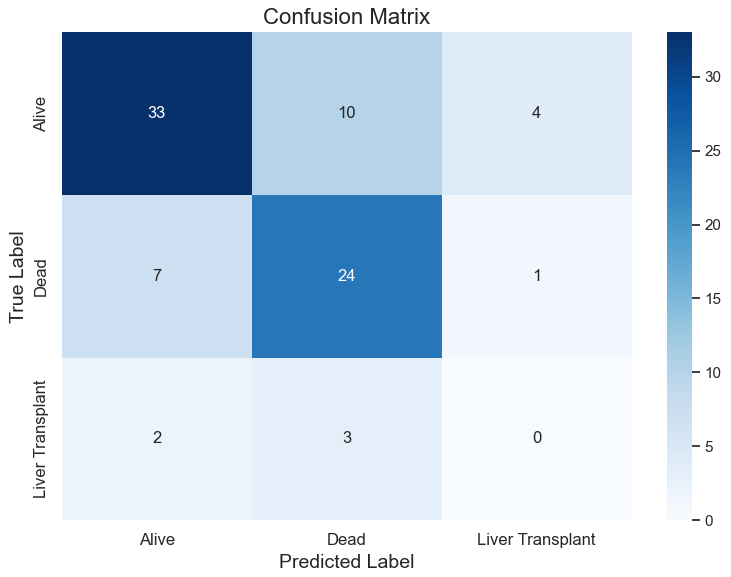

Figure 31: Matriu de confusió del conjunt de test

Quan mirem la matriu de confusió de prova, observem que mentre el model encara mostra una raonable precisió en la classificació de 'Alive' i 'Dead', comet errors substancials en 'Liver Transplant', amb només 3 de 5 correctament classificats, i 2 casos classificats errròniament com 'Alive'.

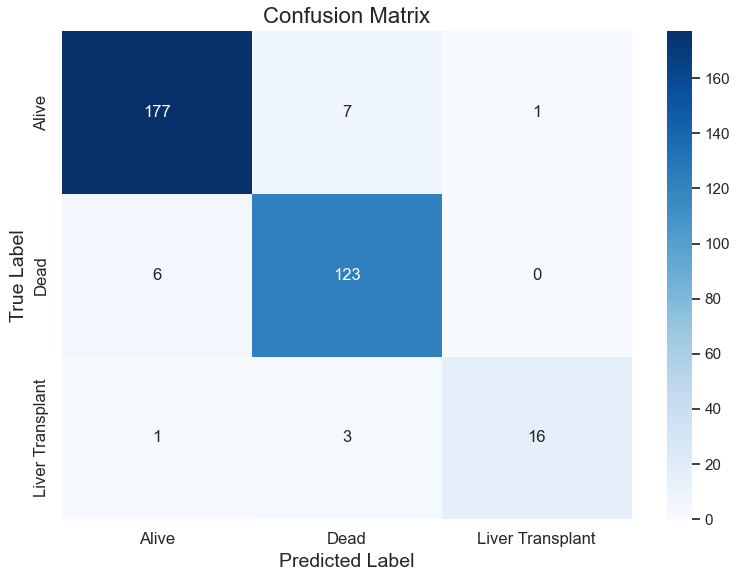

Figure 32: Matriu de confusió del conjunt d'entrenament 'real'

D'altra banda, la matriu de confusió de l'entrenament mostra una alta quantitat de verdaders positius (177 per 'Alive', 123 per 'Dead') i una classificació gairebé perfecta per 'Liver Transplant' (16 de 16). Hi ha poques classificacions errònies, suggerint un excel·lent ajust del model (overfitting).

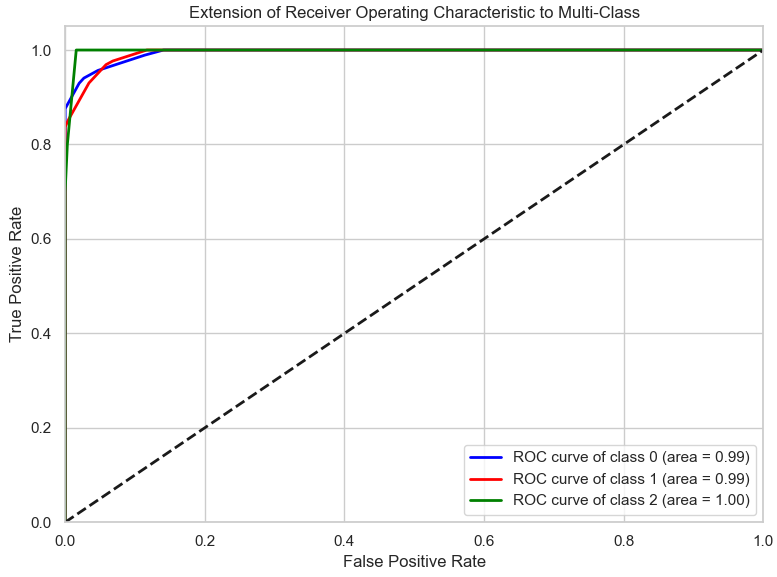

Figure 33: Corba ROC del conjunt d'entrenament

En el conjunt d'entrenament, veiem que l'AUC per a les classes 0 ('Alive') i 1 ('Dead') és de 0.99, i per a la classe 2 ('Liver Transplant') és de 1.00. Aquests valors suggerixen que el model té una capacitat quasi perfecta per distingir entre les classes amb les dades d'entrenament.

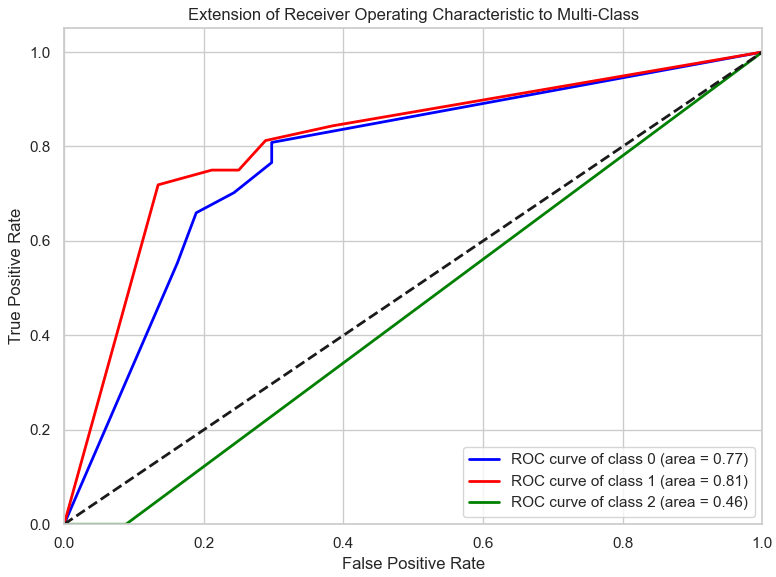

Figure 34: Corba ROC del conjunt de test

En canvi, en el conjunt de proves, l'AUC disminueix a 0.77 per a la classe 0, 0.81 per a la classe 1, i cau dràsticament a 0.46 per a la classe 2. Aquesta reducció indica un decrement en la capacitat predictiva del model en dades no vistes, especialment per a la classe 'Liver Transplant'.

Per concloure, el model d'arbre de decisió demostra signes d'overfitting, com es reflecteix en el contrast entre el rendiment d'entrenament i de prova. L'overfitting es pot atribuir a la naturalesa dels arbres de decisió que tendeixen a crear branques molt específiques per les dades d'entrenament que no necessàriament es generalitzen a noves dades. Per tal de poder evitar-ho es podrien considerar tènciques de poda per restringir la seva profunditat o probar amb un ensemble de models, com ara Random Forest, que pot millorar la generalització mitjançant l'agregació de múltiples arbres.

4.4 Support Vector Machine (SVM)

Les màquines de vectors de suport (SVM) són models poderosos i versàtils, amplament utilitzats en tasques de classificació. Són particularment efectius en espais de dimensions altes i quan la separació entre classes no és clarament definible.

4.4.1 Mètriques de Rendiment

Les mètriques següents s'han triat per a l'avaluació del model SVM:

- **Precisió (Precision)**: Indica la qualitat de les prediccions positives del model.
- **Accuradesa (Accuracy)**: Mesura la proporció total de prediccions correctes.
- **Recall**: Mostra la capacitat del model per identificar tots els casos rellevants.
- **F1-Score**: Proporciona un equilibri entre la precisió i el recall.
- **AUC-ROC**: Evalua la capacitat del model de distingir entre les diverses classes.

4.4.2 Selecció d'Hiperparàmetres

Els hiperparàmetres en un model SVM són crítics i inclouen:

- **C (Penalització)**: El paràmetre de penalització de l'error, que ajuda a controlar l'equilibri entre el marge de decisió suau i la classificació correcta dels punts d'entrenament.
- **kernel**: El tipus de kernel utilitzat per a la transformació de l'espai de característiques. Els més comuns són 'linear', 'poly', 'rbf', i 'sigmoid'.
- **gamma**: El coeficient per a kernels no lineals. Determina la influència d'un únic punt d'entrenament.
- **degree**: El grau del polinomi en el kernel 'poly', si aquest s'escull.

La Taula 24 mostra els hiperparàmetres i els valors que s'han provat.

Hiperparàmetre	**Valors Provats**
C	0.1, 1, 10
kernel	linear, poly, rbf
gamma	scale, auto
degree	2, 3, 4

Table 24: Llista d'hiperparàmetres i valors provats per SVM.

4.4.3 Entrenament i Validació

Hem aplicat una divisió del conjunt de dades del 80% per a entrenament i un 20% per a test, a més d'utilitzar la validació creuada per assegurar la generalitzabilitat del model. Això ens permet ajustar els hiperparàmetres de manera eficaç i identificar si el model pateix de sobreajustament o subajustament.
Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler`
- Encodificar amb `OrdinalEncoder`

4.4.4 Anàlisi dels resultats

Els resultats del model SVM, amb els hiperparàmetres seleccionats, s'han registrat en la següent taula:

Hiperparàmetre	Valor
C	10
degree	2
gamma	scale
kernel	linear

Table 25: Hiperparàmetres optimitzats del model SVM

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts del train en la validació creuada són les següents:

Mètrica	Valor
Accuracy	0.8166
F1 Score	0.8106
Precision	0.8227
Recall	0.8166

Table 26: Mètriques mitjanes d'entrenament en la validació creuada del SVM

Les mètriques mitjanes que ha obtingut aquest model pel que fa als diferents conjunts de validació dins del cross-validation són les següents:

Mètrica	Valor
Accuracy	0.7245
F1 Score	0.7122
Precision	0.7167
Recall	0.7245

Table 27: Mètriques mitjanes de validació en la validació creuada del SVM

El SVM ha mostrat un bon rendiment en el conjunt d'entrenament, però s'observa una disminució en el rendiment en la validació, el que podria indicar una tendència cap a l'overfitting.

Class	Precision	Recall	F1-score	Support
Alive	0.78	0.89	0.83	185
Dead	0.82	0.72	0.77	129
Liver Transplant	0.89	0.40	0.55	20
Accuracy			0.80	334
Macro avg	0.83	0.67	0.72	334
Weighted avg	0.80	0.80	0.79	334

Table 28: Train Classification Report del SVM

La matriu de confusió del conjunt d'entrenament mostra una bona capacitat de classificació per a les classes 'Alive' i 'Dead', però una baixa capacitat per a la classe 'Liver Transplant'.

Class	Precision	Recall	F1-score	Support
Alive	0.80	0.85	0.82	47
Dead	0.73	0.75	0.74	32
Liver Transplant	0.00	0.00	0.00	5
Accuracy			0.76	84
Macro avg	0.51	0.53	0.52	84
Weighted avg	0.72	0.76	0.74	84

Table 29: Test Classification Report del SVM

El rendiment del model en el conjunt de test mostra un descens respecte a l'entrenament, però manté una adequada precisió i F1 score per a 'Alive' i 'Dead', mentre que no es detecta cap cas de 'Liver Transplant'.

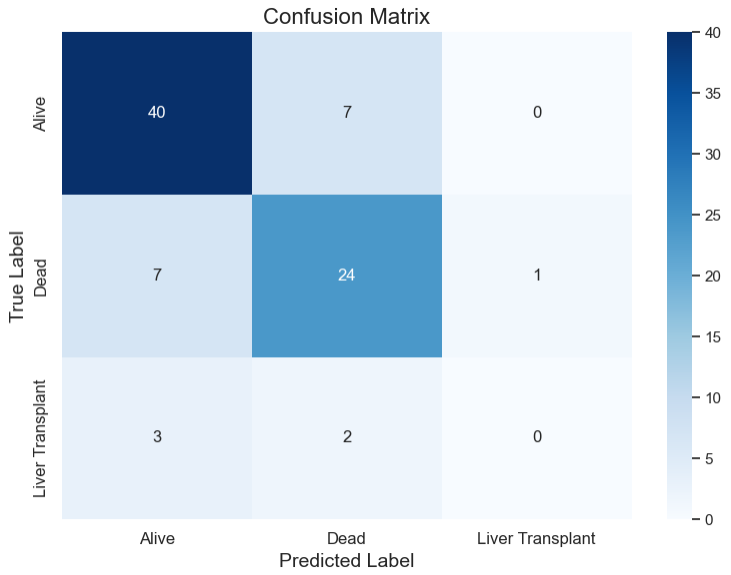

Figure 35: Matriu de confusió del conjunt de test

En el conjunt de proves, el nombre d'errors no és gaire elevat si es compara amb el train. Cal dir però que ha augmentat el nombre d'errors especialment per als pacients 'Dead', on veiem que 7 casos són erròniament classificats com 'Alive'. Això podria ser degut a característiques semblants entre les dues classes que el model SVM no pot separar de manera efectiva.

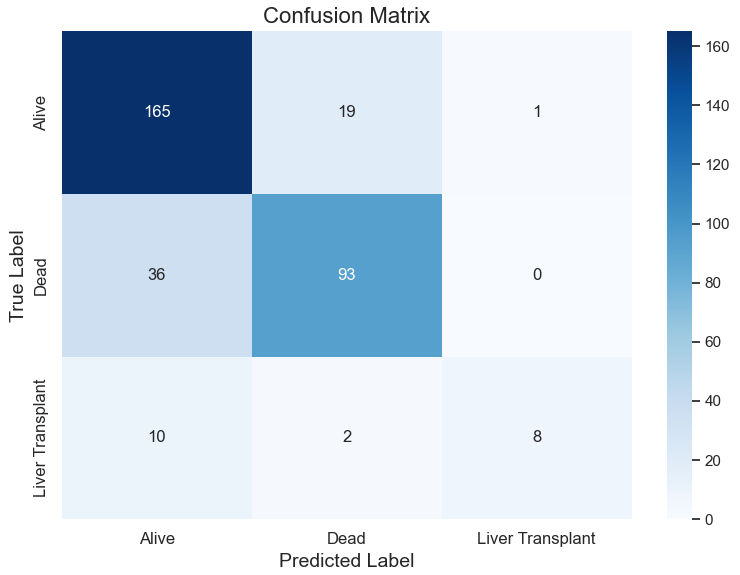

Figure 36: Matriu de confusió del conjunt d'entrenament 'real'

La matriu de confusió per a l'entrenament mostra una certa confusió entre les classes, amb una tendència del model a classificar incorrectament els estats 'Dead' i 'Liver Transplant' com 'Alive'.

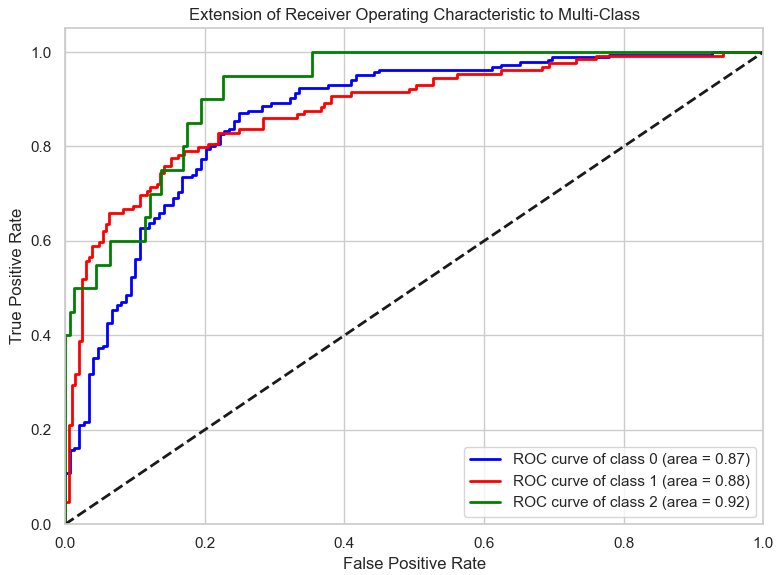

Figure 37: Corba ROC del conjunt d'entrenament

La corba ROC per a l'entrenament mostra AUCs de 0.86 per a 'Alive', 0.88 per a 'Dead', i 0.92 per a 'Liver Transplant', indicant una bona capacitat distintiva del model per a totes les classes. No obstant això, el rendiment en la classe 'Liver Transplant' és més baix comparat amb les altres dues classes.

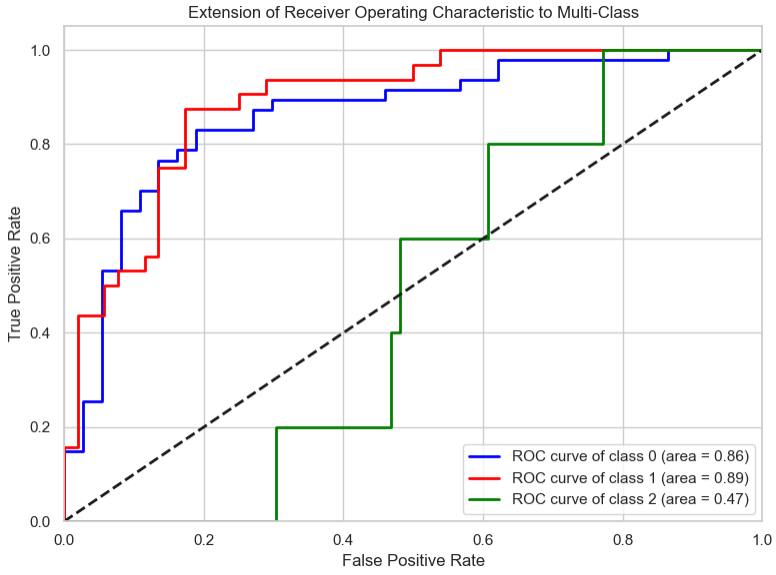

Figure 38: Corba ROC del conjunt de test

Per al conjunt de proves, veiem una disminució en l'AUC de la classe 'Liver Transplant' a 0.49, reflectint una capacitat predictiva pobra per a aquesta classe. Això podria ser a causa de la complexitat de classificar correctament els casos de trasplantament, que podrien ser més raros o menys diferenciats en el conjunt de dades.

El model SVM sembla tenir dificultats per a manejar la classe 'Liver Transplant' tant en l'entrenament com en la prova, com es demostra en les corbes ROC i les matrius de confusió. Això podria ser una indicació que el model necessita un kernel més adequat que pugui capturar millor la frontera de decisió per a aquesta classe particularment complexa. També es podria considerar l'enginyeria de característiques o l'ajust de paràmetres del kernel per millorar la separació entre les classes.

El rendiment respecte a les classes 'Alive' i 'Dead' és més fort, però el model encara comet errors substancials en la predicció d'aquests estats, particularment quan s'apliquen a dades de prova. Això pot indicar que el model pot beneficiar-se d'una millor selecció de característiques o d'una reducció de la dimensionalitat per evitar el soroll i millorar la generalització.

En resum, mentre el model SVM demostra una capacitat decent per classificar 'Alive' i 'Dead', mostra limitacions significatives en la predicció de la classe 'Liver Transplant', suggerint la necessitat de revisar l'estratègia de modelatge i possiblement incorporar coneixements de domini específic per millorar el rendiment en aquesta àrea crítica. Tot i que no predeixi bé la classe 'Liver Trans-

plant' ho fa exactament igual que els altres models observats.

Cal mencionar que també s'han provat diferents tècniques d'oversampling i undersampling alhora per tal de no crear noves classes sintètiques i en tots els casos els resultats eren pitjors que si no feiem cap tècnica de balanceig. A més a més encara que les classes estiguessin equilibrades, el model seguia sense aprendre a classifcar 'Liver Transplant' pel que fa al test. Probablement és degut a que no hi ha cap patró per tal de poder predir si a un pacient la faran un transplant o no.

4.5 Models Extres

En aquesta apartat es mostren dos models nous apart dels demanat en l'enunciat de la pràctica. El motiu era per veure si els resultats es podien millorar ja que amb els models bàsics utilitzats no n'hi havia suficent per a poder crear un model vàlid i acceptat pel que fa al sector de la salut. Degut a que són models extres el seu anàlisi no serà tan rigorós com en els altres tot i que s'analitzarn les mètriques més importants.

4.5.1 Random Forest Classifier

El Random Forest és un mètode d'ensemble que opera construint una multitud d'arbres de decisió en l'entrenament i produint la classe que és el mode de les classes (classificació) o predicció mitjana (regressió) dels arbres individuals. És conegut per la seva robustesa i la seva capacitat de funcionar bé en una àmplia gamma de problemes de classificació.

4.5.1.1 Mètriques de Rendiment Per a l'avaluació del model Random Forest, emprarem les següents mètriques:

- **Precisió (Precision)**: Rellevant quan el cost de falsos positius és elevat.
- **Accuradesa (Accuracy)**: Útil quan les classes estan equilibrades.
- **Recall**: Crític quan és important detectar tots els casos positius.
- **F1-Score**: Útil quan necessitem un balanç entre precisió i recall.
- **AUC-ROC**: Indispensable quan les classes estan desequilibrades.

4.5.1.2 Selecció d'Hiperparàmetres El Random Forest té diversos hiperparàmetres que poden influir significativament en el rendiment del model:

- **n_estimators**: El nombre d'arbres en el bosc.
- **max_depth**: La màxima profunditat dels arbres.
- **min_samples_split**: El nombre mínim de mostres requerides per dividir un node intern.
- **min_samples_leaf**: El nombre mínim de mostres requerides per ser un full de l'arbre.
- **max_features**: El nombre de característiques a considerar quan es busca la millor divisió.
- **bootstrap**: Mètode per a mostrejar les dades utilitzades per construir cada arbre.

La Taula 30 presenta una llista dels hiperparàmetres considerats i els valors que s'han provat.

Hiperparàmetre	**Valors Provats**
n_estimators	100, 200, 500
max_depth	10, 20, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	sqrt, log2, None
bootstrap	True, False

Table 30: Llista d'hiperparàmetres i valors provats per Random Forest.

4.5.1.3 Entrenament i Validació Hem aplicat una divisió del conjunt de dades del 80% per a entrenament i un 20% per a test, a més d'utilitzar la validació creuada per assegurar la generalitzabilitat del model. Això ens permet ajustar els hiperparàmetres de manera eficaç i identificar si el model pateix de sobreajustament o subajustament.
Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler`
- Encodificar amb `OrdinalEncoder`

Els millors hiperparàmetres per aquest model han estat els següents:

Hiperparàmetre	**Valor**
bootstrap	True
max_depth	None
max_features	sqrt
min_samples_leaf	1
min_samples_split	5
n_estimators	100

Table 31: Millors hiperparàmetres utilitzats en el model Random Forest.

Estat	**Precisió**	**Recuperació (Recall)**	**Puntuació F1**	**Suport**
Alive	0.87	0.87	0.87	47
Dead	0.76	0.88	0.81	32
Liver Transplant	0.00	0.00	0.00	5
Precisió mitjana		0.82		84

Table 32: Informe de classificació per a la partició de prova del model Random Forest

Com es pot observar el `Random Forest` ens ha donat molts millors resultats que els altres models que s'han vist anteriorment.

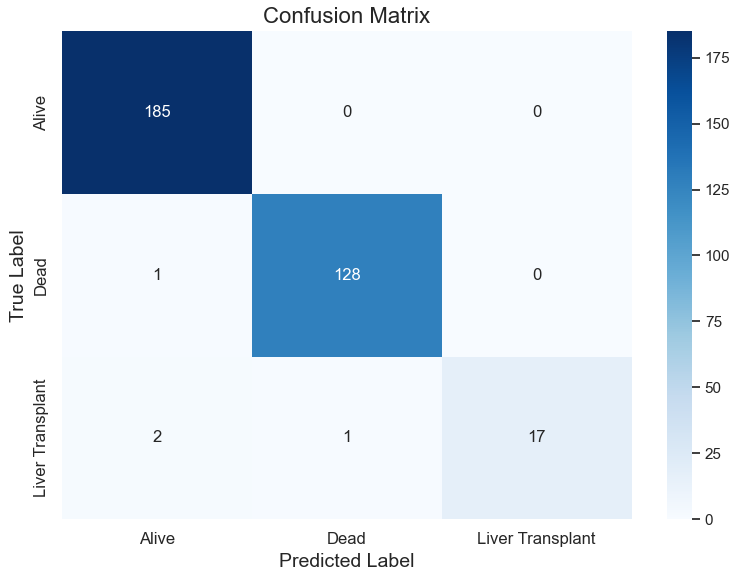

Figure 39: Confusion Matrix Train

Finalment cal dir que el model té un overfitting gairebé total com es pot observar en la imatge 39. Per tant, potser caldria ajustar millor la profunditat de les branques per evitar aquesta gairebé perfecció en les dades d'entrenament.

4.5.2 XGBoost Classifier

XGBoost (eXtreme Gradient Boosting) és un algorisme d'aprenentatge supervisat que implementa un procés de boosting de gradient, el qual és altament eficient en termes de velocitat i rendiment. XGBoost és ampliament reconegut per la seva capacitat de manejar gran varietat de tipus de dades, la seva robustesa en la presència de dades anòmales i la seva eficiència en grans conjunts de dades.

4.5.2.1 Mètriques de Rendiment Les mètriques seleccionades per avaluar el rendiment del model XGBoost inclouen:

- **Precisió (Precision)**: Útil quan el cost de falsos positius és considerable.
- **Accuradesa (Accuracy)**: Mesura general de rendiment quan les classes estan equilibrades.
- **Recall**: Important quan és crític detectar tots els casos positius.
- **F1-Score**: Combina precisió i recall en una sola mètrica per a casos on és necessari un balanç.
- **AUC-ROC**: Avaluació de la capacitat discriminativa del model, essencial en classes desequilibrades.

4.5.2.2 Selecció d'Hiperparàmetres Els hiperparàmetres clau en XGBoost que necessiten ser afinats inclouen:

- **eta (learning rate)**: Controla la contribució de cada arbre en el model final.
- **max_depth**: Determina la profunditat màxima de cada arbre.
- **subsample**: Fracció de mostres a utilitzar per construir cada arbre, per prevenir sobreajustament.
- **colsample_bytree**: Fracció de característiques a utilitzar per construir cada arbre.
- **n_estimators**: Nombre d'arbres a construir.
- **objective**: La funció d'objectiu utilitzada per la tasca de predicció.

Els valors específics provats per aquests hiperparàmetres es mostren a la Taula 33.

Hiperparàmetre	**Valors Provats**
eta (learning rate)	0.01, 0.1, 0.3
max_depth	3, 6, 9
subsample	0.5, 0.7, 1.0
colsample_bytree	0.5, 0.7, 1.0
n_estimators	100, 200, 300
gamma	0, 1, 5

Table 33: Llista d'hiperparàmetres i valors provats per XGBoost.

4.5.2.3 Entrenament i Validació Hem aplicat una divisió del conjunt de dades del 80% per a entrenament i un 20% per a test, a més d'utilitzar la validació creuada per assegurar la generalitzabilitat del model. Això ens permet ajustar els hiperparàmetres de manera eficaç i identificar si el model pateix de sobreajustament o subajustament.
Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler`
- Encodificar amb `OrdinalEncoder`

Els millors hipeparàmetres del model han estat els següents:

Classe	**Precisió**	**Recuperació (Recall)**	**Puntuació F1**	**Suport**
0	0.82	0.85	0.83	47
1	0.71	0.78	0.75	32
2	0.00	0.00	0.00	5
Precisió mitjana		0.77		84

Table 35: Informe de classificació per a la partició de prova del model XGBoost.

Hiperparàmetre	Valor
colsample_bytree	1.0
gamma	0
learning_rate	0.1
max_depth	9
n_estimators	300
subsample	0.7

Table 34: Hiperparàmetres utilitzats en el model XGBoost.

Com es pot veure, els resultats del model són lleugerament inferiors als del `Random Forest` però notablement superiors als que s'han probat anteriorment.

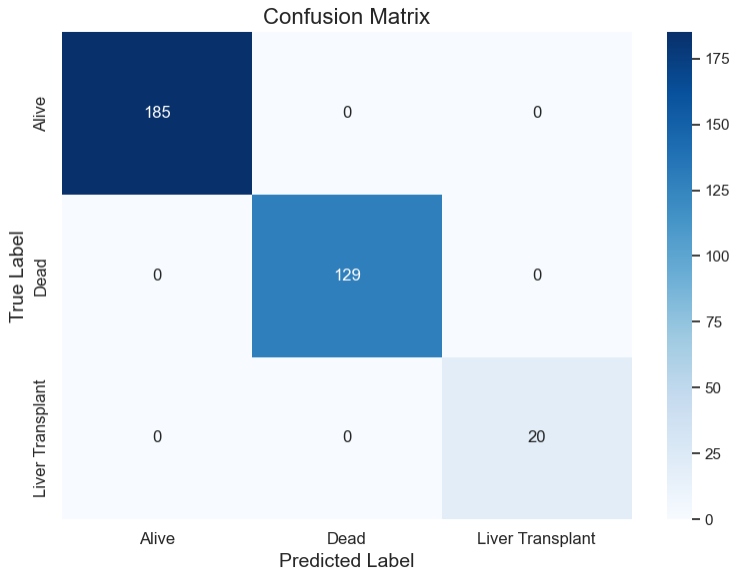

Figure 40: Confusion Matrix Train

En la matriu de confusió del train (figura 40) es pot veure un overfitting total de 100%, cosa que ens indica que aquest model necessita unes modificacions.

4.6 Selecció del Model

Cal mencionar que encara que el model escollit no sigui el millor dels que s'han provat degut a que s'obtenen molts millors resultats amb un `Random Forest` o amb una `Explainable Boosting Machine` com que aquest no formaven part de la part obligatòria del treball no els tindrem presents alhora de la selecció.

Finalment s'ha seleccionat un model SVM per a la classificació dels estats de salut dels pacients en les classes:

- `Alive`
- `Dead`
- `Liver Transplant`

Amb els següents hiperparàmetres:

Hiperparàmetre	**Valor**
C	10
degree	2
gamma	scale
kernel	linear

Table 36: Hiperparàmetres optimitzats del model SVM

A continuació es detallen les seves capacitats, limitacions i rendiment en les diferents particions del conjunt de dades.

Descripció del Model Triat

El SVM seleccionat amb un kernel lineal i els hiperparàmetres ajustats ha demostrat ser capaç de capturar la frontera de decisió entre les classes. No obstant això, els resultats varien significativament entre el conjunt d'entrenament i de prova, suggerint qüestions de generalització que s'han d'abordar.

Capacitats del Model

El model ha mostrat una bona capacitat de classificació en el conjunt d'entrenament amb una precisió general (*accuracy*) de 0.83 i una puntuació F1 (*f1-score*) de 0.829. Això indica que el model pot separar les classes amb eficàcia en les dades conegudes.

Limitacions del Model

Tot i que el model SVM ha proporcionat un rendiment decent en l'entrenament, les seves limitacions es fan evidents en la partició de prova. Amb una precisió de 0.726 i una puntuació F1 de 0.750, hi ha una caiguda notable respecte als resultats de l'entrenament. Això és particularment crític en l'àmbit de la salut, on la capacitat de predir correctament els estats de salut dels pacients és d'importància vital. El model no ha pogut classificar cap pacient en la categoria 'Liver Transplant' correctament, el que ressalta la necessitat d'un model més robust i precis. Però com ja s'ha anat mencionant al llarg de l'informe és possible que el transplantament de fetge no es pugui predir ni trobar en patrons degut a que podria ser un simple fet aleatori així que en l'evaluació no es tindrà en compte que no sàpiga classificar aquesta classe; a més a més en l'àmbit de salut és més important saber si un pacient sobreviurà o no per tal de poder prendre les decisions adequades.

Resultats en la Partició de Test

Els resultats en la partició de test són inferiors als del conjunt d'entrenament i validació. Això pot ser indicatiu de l'overfitting al conjunt d'entrenament, o bé d'un conjunt de proves que presenta característiques significativament diferents que no han estat apreses pel model.

Conclusions

Tot i que el model SVM mostra potencial, les seves limitacions en la generalització i la sensibilitat als desequilibris de les classes fan que no sigui suficientment fiable per a aplicacions crítiques com la medicina. Es recomana explorar altres models o tècniques d'ensemble, així com realitzar una anàlisi més profunda de les dades i potser incorporar més informació de domini per millorar la precisió de les prediccions.

Taula de Resultats

A continuació, es presenten els resultats de l'avaluació del model:

Metric	Train	Validation	Test
Accuracy	0.835	0.616	0.726
F1 Score	0.835	0.645	0.750
Precision	0.835	0.704	0.790
Recall	0.835	0.616	0.730

Table 37: Comparació de resultats entre les particions d'entrenament, validació i prova.

4.7 Model Card

Model Card: SVM per a la Predicció de l'Estat de Salut dels Pacients December 28, 2023

Model Details

- **Developing Organization:** Roger Baiges Trilla, `Facultat Informàtic de Barcelona`
- **Model Date:** December 28, 2023
- **Model Version:** 1.0
- **Model Type:** Support Vector Machine (SVM)
- **Training Algorithms:** SVM amb kernel lineal
- **Hyperparameters:** $C = 10$, degree = 2, gamma = scale
- **License:** MIT License
- **Contact Information:** [roger.baiges.trilla@estudiantat.upc.edu]

Intended Use

- **Primary Intended Uses:** Aquest model està destinat a ser una eina de suport en la presa de decisions clíniques, proporcionant una segona opinió sobre l'estat de salut dels pacients.
- **Primary Intended Users:** Professionals de la salut en hospitals i clíniques.
- **Out-of-Scope Use Cases:** No és apropiat utilitzar aquest model com a única font de diagnòstic o per a decisions de tractament sense supervisió mèdica.

Factors

- **Relevant Factors:** Dades demogràfiques dels pacients, biomarcadors clínics, història mèdica.
- **Evaluation Factors:** Precisió del model, sensibilitat, especificitat, puntuació F1.

Metrics

- **Model Performance Measures:** Veure secció de Rendiment a continuació.
- **Decision Thresholds:** No aplicable, ja que el model SVM utilitza màrgens.
- **Variation Approaches:** Validació creuada.

Evaluation Data

- **Datasets:** Conjunt de dades anònimes de pacients amb cirrosis.
- **Motivation:** Millorar la predicció de l'estat de salut i assistència als pacients.
- **Preprocessing:** Normalització de les característiques, tractament de valors mancants, codificació de característiques categòriques.

Training Data

El conjunt de dades d'entrenament reflecteix la distribució de la població de pacients amb condicions hepàtiques, incloent diverses edats, gèneres i etapes de la malaltia.

Quantitative Analyses

- **Unitary Results:** AUC, precisió, puntuació F1, com s'ha indicat anteriorment.
- **Intersectional Results:** Anàlisi de la capacitat del model per a diferents subgrups de població no ha estat realitzada.

Ethical Considerations

L'ús d'aquest model ha de ser acompanyat per la comprensió que, malgrat les seves prometedores mètriques de rendiment, no pot reemplaçar el judici clínic humà i ha de ser utilitzat amb precaució.

Caveats and Recommendations

El model no ha demostrat ser suficientment fiable per a ser utilitzat com a única eina diagnòstica. Es recomana la seva utilització com a part d'un conjunt més ampli d'eines de diagnòstic, incloent avaluació mèdica i proves addicionals.

Rendiment

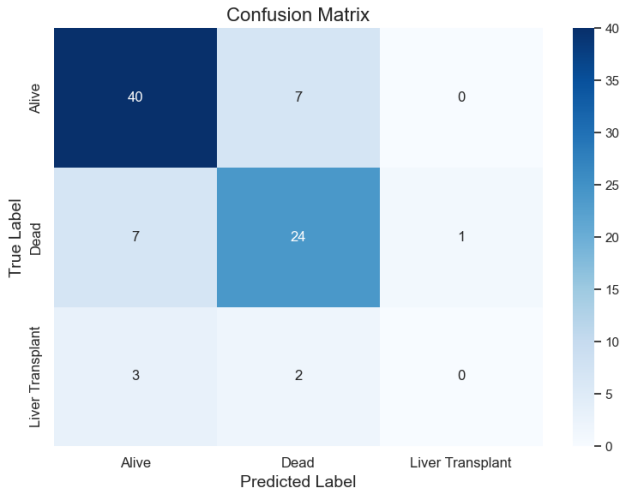

Figure 41: Matriu de Confusió per a la Partició de Test

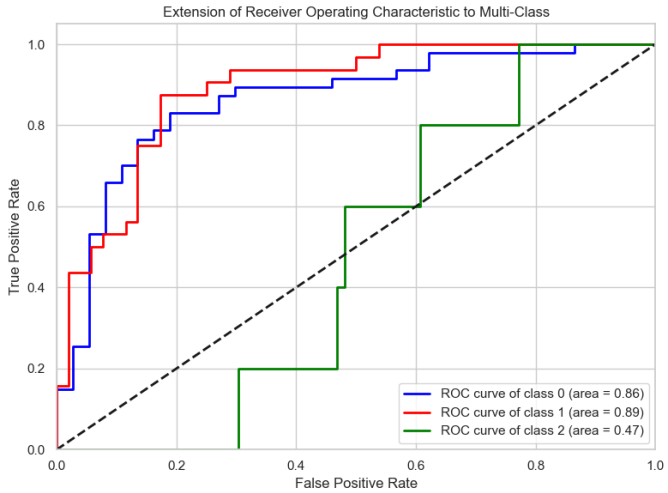

Figure 42: Corba ROC per a la Partició de Test

5 Bonus

A continuació es presenten els apartats *Bonus* del projecte. Sent el primer un model `EBM` i el segon un `clustering`.

5.1 Explainable Boosting Machine (EBM)

Explainable Boosting Machine (EBM) és un model de màquina d'aprenentatge supervisat que combina la potència predictiva dels mètodes de màquines d'aprenentatge amb la interpretabilitat dels models lineals. EBM utilitza boosting en un conjunt de models simples per millorar la precisió mantenint la comprensibilitat.

5.1.1 Mètriques de Rendiment

Les següents mètriques s'utilitzen per avaluar el rendiment de l'EBM:

- **Precisió (Precision)**: Indica la qualitat de les prediccions positives del model, útil en situacions on els falsos positius són costosos.
- **Accuradesa (Accuracy)**: Mesura la proporció total de prediccions correctes, important quan les classes són equilibrades.
- **Recall**: És essencial quan és important identificar tots els casos positius, com en situacions mèdiques.
- **F1-Score**: Harmonitza la precisió i el recall, útil quan es necessita un balanç entre aquestes mètriques.
- **AUC-ROC**: Avalua la capacitat del model de distingir entre classes, crucial per a classes desequilibrades.

5.1.2 Selecció d'Hiperparàmetres

Els hiperparàmetres en un EBM poden incloure:

- **learning_rate**: La taxa d'aprenentatge per a l'optimització, controla la velocitat a la qual el model aprèn.
- **max_bins**: El nombre màxim de bins que s'utilitzen per discretitzar les variables contínues.
- **max_leaves**: El nombre màxim de fulles per arbre de decisió, que defineix la complexitat del model.
- **min_samples_leaf**: El nombre mínim de mostres permeses en una fulla de l'arbre.

Els valors provats per aquests hiperparàmetres poden ser visualitzats en la Taula 38.

Hiperparàmetre	Valors Provats
learning_rate	0.01, 0.1, 0.2
max_bins	256, 512
max_leaves	3, 5, 10
min_samples_leaf	1, 5, 10

Table 38: Llista d'hiperparàmetres i valors provats per EBM.

5.1.3 Entrenament i Validació

Hem aplicat una divisió del conjunt de dades del 80% per a entrenament i un 20% per a test, a més d'utilitzar la validació creuada per assegurar la generalitzabilitat del model. Això ens permet ajustar els hiperparàmetres de manera eficaç i identificar si el model pateix de sobreajustament o subajustament.
Els tractaments que s'han realitzat a les dades han estat:

- Imputar els valors faltants
- Normalitzar fent ús de la funció `StandardScaler` (encara que no afectava en els resultats)
- Encodificar amb `OrdinalEncoder`

5.1.4 Anàlisi dels resultats de l'Explainable Boosting Machine

L'Explainable Boosting Machine (EBM) ha demostrat ser extremadament eficaç en el conjunt d'entrenament, assolint una precisió gairebé perfecta. A continuació es detallen les mètriques específiques aconseguides:

Class	Precision	Recall	F1-score	Support
Alive	0.98	1.00	0.99	185
Dead	0.99	0.99	0.99	129
Liver Transplant	1.00	0.85	0.92	20
Accuracy			0.99	334
Macro avg	0.99	0.95	0.97	334
Weighted avg	0.99	0.99	0.99	334

Table 39: Train Classification Report de l'Explainable Boosting Machine

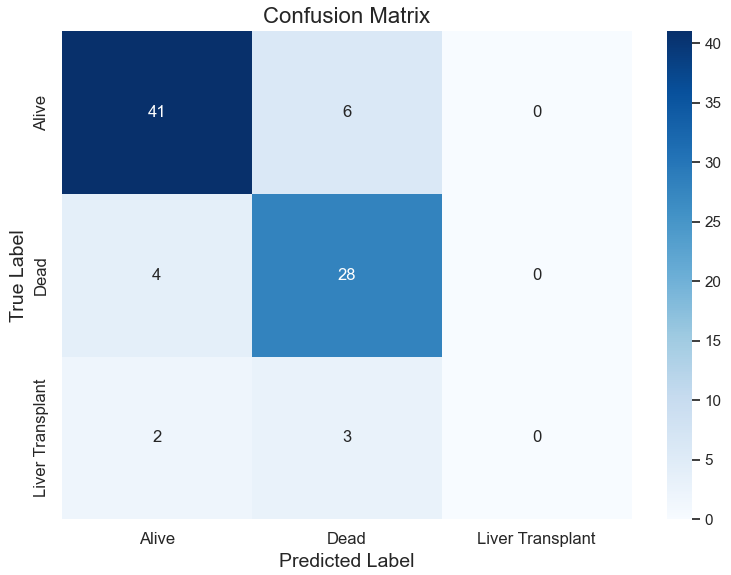

Figure 43: Matriu de confusió del train

La matriu de confusió del conjunt d'entrenament revela que l'EBM ha classificat correctament tots els casos de 'Alive', amb una precisió lleugerament inferior en les altres dues classes, encara que segueix sent molt alta.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Alive	0.87	0.87	0.87	47
Dead	0.76	0.88	0.81	32
Liver Transplant	0.00	0.00	0.00	5
Accuracy			0.82	84
Macro avg	0.54	0.58	0.56	84
Weighted avg	0.78	0.82	0.80	84

Table 40: Test Classification Report de l'Explainable Boosting Machine

Pel que fa al test podem observar que també és un dels models més poderosos i tan sols es pot comparar amb el `Random Forest`. Encara que obviament el rendiment és més baix que en el test i que no ha predit bé cap transplantament de fetge.

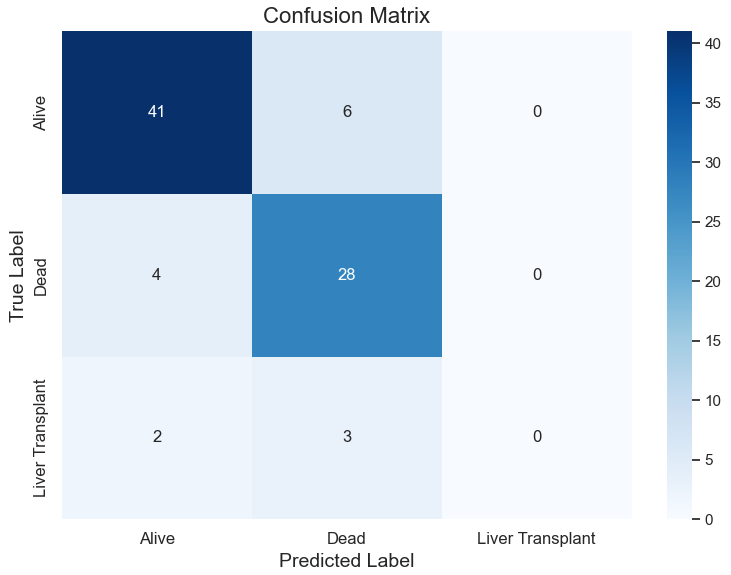

Figure 44: Matriu de confusió del test

La matriu de confusió del conjunt de test mostra que, malgrat una lleugera disminució de la precisió en comparació amb el conjunt d'entrenament, l'EBM manté un rendiment robust. No obstant això, cal destacar que la classe 'Liver Transplant' no ha estat correctament identificada en cap ocasió, la qual cosa indica que el model pot necessitar més dades o una millor representació d'aquesta classe per aprendre a classificar-la correctament.

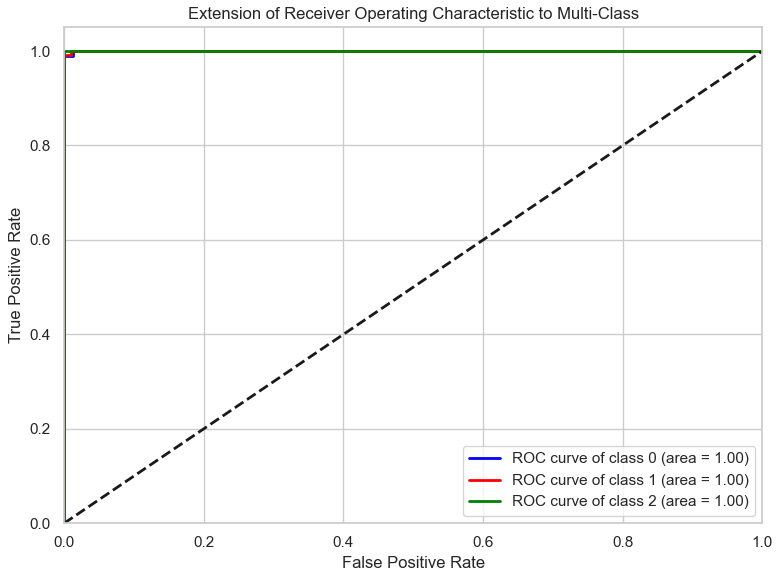

Figure 45: Corba ROC del conjunt de train

Com es pot veure en la figura 45 els resultats són perfectes indicant així un overfitting brutal.

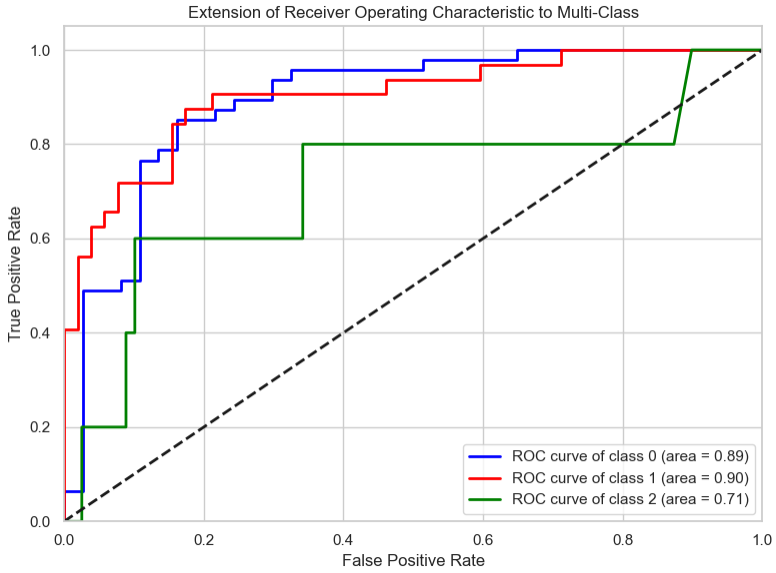

Figure 46: Corba ROC del conjunt de test

En aquesta corba ROC podem veure que el model també ho fa bastant bé encara que, obviament,

es nota que les dades han fet un overfitting total. Destaquem la predicció de 'Liver Transplant' que és molt millor que en els altres models.

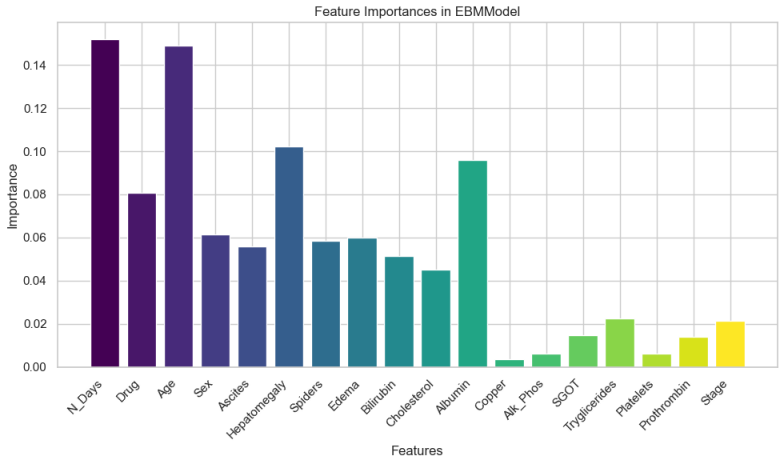

Figure 47: Imatges de les features més importants

Com es pot veure en la figura 47 les variables més importants són el nombre de dies que porta un pacient en l'experiment; tot i que aquesta variable podria no ser útil si volguessim aplicar el model a pacients desconeguts que no han participat en l'experiment. Seguidament per la variable `Age` i per `Hepatomegaly`. Les variables menys importants són `Copper`, `Alk_Phos` i `Platelets`.

Finalment, aquests resultats suggereixen que mentre que l'EBM té un rendiment excepcional en el conjunt d'entrenament, hi ha marge de millora en la seva capacitat de generalització, especialment en el que respecta a les classes minoritàries com 'Liver Transplant'. La seva capacitat explicativa podria ser utilitzada per aprofundir en les característiques que porten a aquestes confusions i ajudar a millorar la interpretació dels resultats clínics.

5.2 Clustering

L'anàlisi de clústers s'ha realitzat per identificar patrons dins de les dades i agrupar els subjectes en grups amb característiques similars. S'ha utilitzat un mètode de clústering que ha resultat en quatre grups distintius, cadascun amb les seves pròpies estadístiques descriptives per a les variables numèriques i categòriques.

Per tal de poder decidir quin era el nombre adequat de clústers es va realitzar un dendrograma, com es pot veure en la figura 48, on es pot veure que el nombre adequat es troba entre 3 i 4. Això és degut a que podem trobar molta distància entre els grups cosa que no passaria si el nombre de clústers fos més gran. Finalment es va decidir que el nombre de grups a analitzar seria 4 encara que un d'ells és molt petit per tal de veure si realment es noten les diferències.

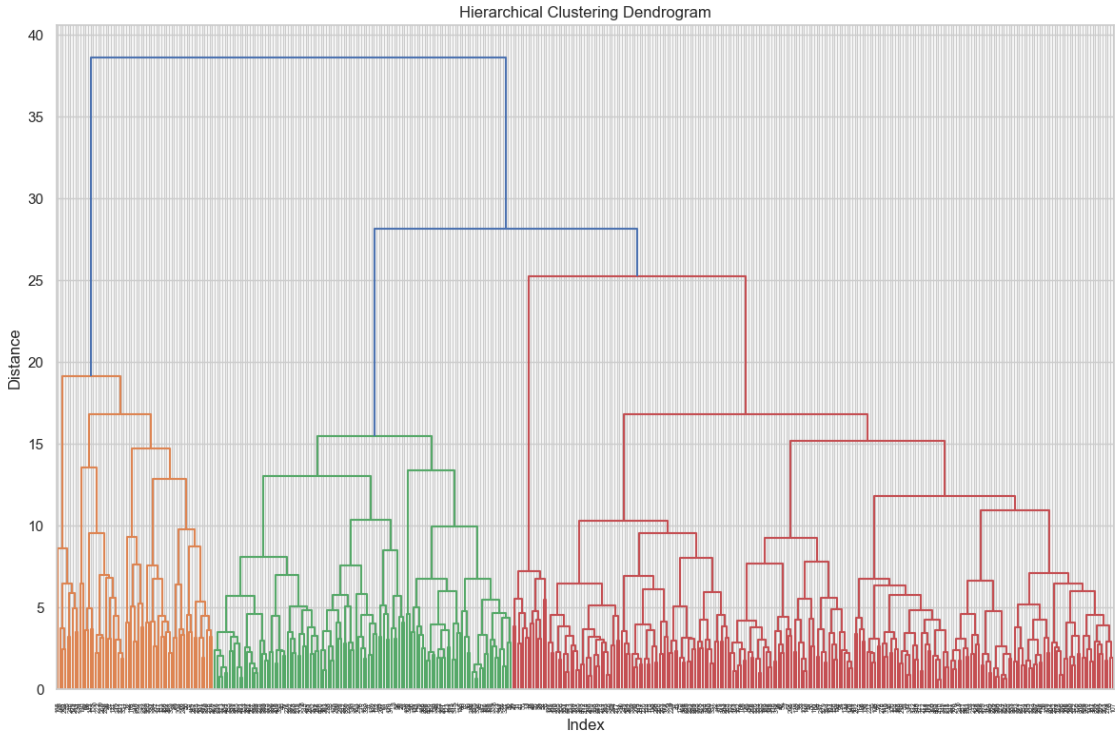

Figure 48: Dendrograma de les dades

En la figura 49 podem observar la projecció dels diferents clústers sobre els dos primers components de l'ACP. Com es pot veure hi ha una clara diferència entre els clústers 1, 2 i 4 però no tant en el grup 3; que justament és el més petit. Això probablement significa que amb tan sols dos components no n'hi ha prou per poder veure la diferència del tercer clúster però si la dels altres. Molt probablement les diferències es trobin també en variables categòriques i, per tant, no sigui suficient amb un anàlisi de components principals.

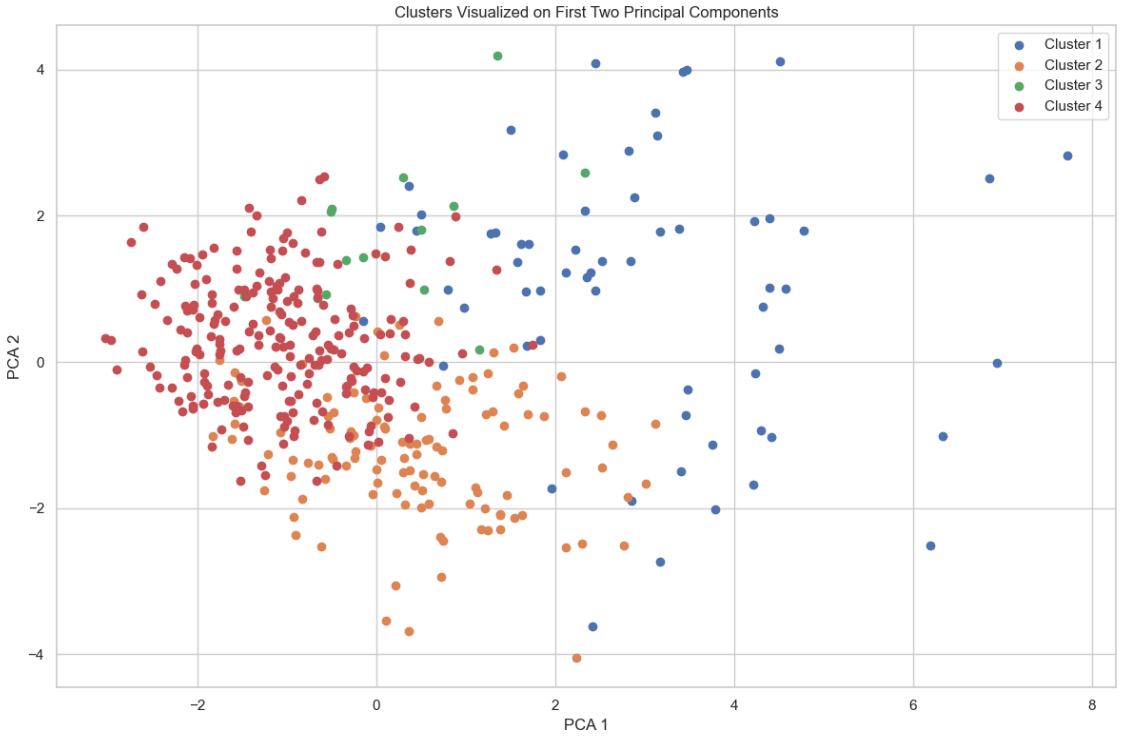

Figure 49: Imatge sobre els clusters projectats en els dos primers components de l'ACP

5.2.1 Perfil del Clúster 1

Aquest grup inclou un total de 62 pacients amb valors significativament alts de bilirrubina (mitjana de 10.91), indicant una possible severitat més gran en la seva condició hepàtica en comparació amb els altres grups. El colesterol també és notablement alt (mitjana de 628), el que podria ser indicatiu de complicacions addicionals relacionades amb el fetge. Aquest grup presenta una majoria de dones (85.48%) i una inclinació cap a l'ús de D-penicil·lamina com a tractament. També s'observa una propensió a tenir hepatomegàlia (82.58%) i aranyes vasculars (51.61%). La majoria dels pacients es troben en l'etapa 4 de la malaltia.

Variable	Mitjana	Desviació Estàndard	Mínim	Màxim
N_Days	1193.79	964.85	41.00	4191.00
Age	18425.68	3522.91	11273.00	28650.00
Bilirubin	10.91	6.51	0.90	28.00
Cholesterol	628.10	383.39	175.00	1775.00
Albumin	3.28	0.49	2.10	4.16
Copper	187.04	118.89	34.00	588.00
Alk_Phos	2309.20	1240.08	559.00	6064.80
Tryglicerides	194.78	84.82	55.00	598.00
Platelets	281.16	121.43	62.00	721.00
Prothrombin	11.10	1.12	9.50	15.20

Table 41: Estadístiques de les variables numèriques per al Clúster 1

Variable	Moda	Freqüència Moda	%
Drug	D-penicil·lamina	32	0.516129
Sex	F	53	0.854839
Ascites	N	48	0.774194
Hepatomegàlia	Y	51	0.822581
Aranyes vasculars	Y	32	0.516129
Edema	N	42	0.677419
Etapa	4.0	28.0	0.451613

Table 42: Estadístiques de les variables categòriques per al Clúster 1

5.2.2 Perfil del Clúster 2

Els 118 pacients d'aquest grup mostren nivells més moderats de bilirrubina (mitjana de 2.74) i un perfil de colesterol considerablement més baix (mitjana de 316.87) en comparació amb el Clúster 1. Aquest grup també està dominat per dones (81.36%) i majoritàriament tractades amb D-penicil·lamina. Es destaca la prevalença d'ascitis (90.68%) i hepatomegàlia (85.59%), però amb una menor incidència d'aranyes vasculars (69.49%). Igual que en el clúster anterior, la majoria de pacients es troben en l'etapa 4 de la cirrosis.

Variable	Mitjana	Desviació Estàndard	Mínim	Màxim
N_Days	1474.02	1067.51	41.00	4795.00
Age	20917.33	3342.25	12285.00	28018.00
Bilirubin	2.74	2.44	0.40	15.00
Cholesterol	316.87	93.12	151.00	636.00
Albumin	3.28	0.41	1.96	4.19
Copper	103.27	60.43	10.00	290.00
Alk_Phos	1731.50	1201.98	516.00	7277.00
SGOT	120.49	37.12	43.40	210.80
Tryglicerides	106.69	24.36	49.00	188.00
Platelets	206.84	75.69	71.00	410.00
Prothrombin	11.28	1.29	9.50	18.00

Table 43: Estadístiques de les variables numèriques per al Clúster 2

Variable	Moda	Freqüència Moda	%
Drug	D-penicil·lamina	83	0.70339
Sex	F	96	0.813559
Ascites	N	107	0.90678
Hepatomegàlia	Y	101	0.855932
Aranyes vasculars	N	82	0.694915
Edema	N	87	0.737288
Etapa	4.0	78.0	0.661017

Table 44: Estadístiques de les variables categòriques per al Clúster 2

5.2.3 Perfil del Clúster 3

El Clúster 3 presenta el nombre més petit de subjectes, concretament 14, però amb la durada més llarga en dies, que podria indicar un estudi longitudinal o seguiment a llarg termini d'aquests pacients. Aquest grup té nivells moderats de bilirrubina i colesterol, i els més alts nivells d'albumina, suggerint millor funció hepàtica.

Està caracteritzat per tenir els valors més alts en l'Alk_Phos (mitjana de 9977.44), suggerint alteracions significatives en la funció hepàtica. Els nivells de bilirrubina són moderats (mitjana de 1.95) i el colesterol és comparable al Clúster 2 (mitjana de 324.99). Predominen les dones (85.71%), i cap d'ells pateix d'ascitis. La majoria dels pacients en aquest grup estan en l'etapa 3 de la malaltia.

Variable	Mitjana	Desviació Estàndard	Mínim	Màxim
N_Days	3279.00	1129.88	1360.00	4523.00
Age	17983.29	3629.13	12279.00	24020.00
Bilirubin	1.95	1.48	0.70	5.70
Cholesterol	324.99	82.38	231.00	498.00
Albumin	3.49	0.37	2.84	4.14
Copper	132.36	69.52	49.00	281.00
Alk_Phos	9977.44	1934.44	7277.00	13862.40
SGOT	112.05	43.22	56.76	206.40
Tryglicerides	148.17	61.57	88.00	319.00
Platelets	324.29	116.05	195.00	563.00
Prothrombin	11.00	0.70	10.30	12.70

Table 45: Estadístiques de les variables numèriques per al Clúster 3

Variable	Moda	Freqüència Moda	%
Drug	D-penicil·lamina	8	0.571429
Sex	F	12	0.857143
Ascites	N	14	1.0
Hepatomegàlia	Y	9	0.642857
Aranyes vasculars	N	10	0.714286
Edema	N	13	0.928571
Etapa	3.0	8.0	0.571429

Table 46: Estadístiques de les variables categòriques per al Clúster 3

5.2.4 Perfil del Clúster 4

Finalment aquest últim grup, a banda de ser el clúster amb més població amb un total de 224 malalts, presenta els valors més baixos de bilirrubina (mitjana de 1.42) i Copper (mitjana de 65.52), suggerint una menor gravetat en termes de disfunció hepàtica. El colesterol i l'Alk_Phos són moderadament alts (mitjana de 320.11 i 1426.54, respectivament). La majoria dels pacients no presenten ascitis, hepatomegàlia ni aranyes vasculars.

Variable	Mitjana	Desviació Estàndard	Mínim	Màxim
N_Days	2266.87	930.74	198.00	4556.00
Age	17341.69	3570.47	9598.00	27398.00
Bilirubin	1.43	1.25	0.30	8.90
Cholesterol	320.11	87.56	120.00	660.00
Albumin	3.67	0.33	2.48	4.64
Copper	65.52	39.34	4.00	227.00
Alk_Phos	1426.54	900.13	289.00	5890.00
SGOT	109.61	41.21	26.35	246.45
Tryglicerides	111.29	38.08	33.00	260.00
Platelets	271.92	88.43	92.00	539.00
Prothrombin	10.33	0.58	9.00	12.00

Table 47: Estadístiques de les variables numèriques per al Clúster 4

Variable	Moda	Freqüència Moda	%
Drug	Placebo	116	0.517857
Sex	F	213	0.950893
Ascites	N	223	0.995536
Hepatomegàlia	N	151	0.674107
Aranyes vasculars	N	195	0.870536
Edema	N	212	0.946429
Etapa	3.0	102.0	0.455357

Table 48: Estadístiques de les variables categòriques per al Clúster 4

5.2.5 Conclusions del clustering

El clustering ha permès visualitzar diferents perfils de pacients, oferint una comprensió més profunda de les característiques clíniques que poden influir en la seva evolució. Aquesta segmentació dels pacients en grups homogenis és crucial per a la nostra tasca de predicció, ja que:

- **Personalització de Prediccions**: Els clústers proporcionen una base per a modelar més precisament les trajectòries de la malaltia. Per exemple, pacients amb alts nivells de bilirrubina i colesterol (Clúster 1) poden tenir un risc més elevat de resultats adversos, el que es pot integrar en els models predictius.
- **Identificació de Factors de Risc**: Els clústers han revelat factors de risc específics i característiques clíniques significatives que poden ser indicadors clau en la predicció de la necessitat de trasplantament, supervivència o mort.
- **Enfocament en Grups Específics**: Els clústers ens permeten centrar-nos en grups de pacients amb necessitats especials, com els del Clúster 3, amb alts nivells d'Alk_Phos, que podrien requerir seguiment més intensiu.

Fora de la nostra tasca directa de predicció, els resultats del clustering tenen aplicacions potencials com:

- **Suport a Decisions Clíniques**: La classificació dels pacients en clústers pot ajudar els metges a prendre decisions més informades sobre tractaments i maneig de pacients.
- **Recerca Biomèdica**: Els resultats poden ser útils en investigacions biomèdiques per explorar com les diferents característiques clíniques interaccionen i afecten la progressió de la cirrosi.
- **Polítiques de Salut Pública**: Aquesta segmentació pot servir per a la planificació de recursos sanitaris, especialment en l'assignació de fons per a programes de prevenció i tractament de la cirrosi.

El clustering realitzat ofereix una perspectiva valuosa no només per a la nostra tasca de predicció, sinó també per a la comprensió general de la cirrosi i la seva gestió. Aquesta metodologia ens permet personalitzar millor els models predictius i pot ser una eina valuosa en diversos àmbits relacionats amb la salut i la recerca.

6 Conclusions

Al llarg d'aquest projecte, s'ha realitzat una anàlisi exhaustiva del dataset *Cirrhosis Patient Survival Prediction*, amb l'objectiu d'entendre millor les dinàmiques i les característiques dels pacients afectats per cirrosi.

Anàlisi Exploratòria de les Dades

S'ha iniciat el projecte amb una anàlisi exploratòria de les dades, on s'han examinat les distribucions de les diferents característiques. Aquesta anàlisi ha revelat patrons significatius i anomalies que han estat crucials per a l'enteniment del dataset i per a la posterior aplicació de models predictius.

Gestió de Dades Perdudes

Un dels reptes claus ha estat el tractament de valors perduts. Per a afrontar aquesta qüestió, s'ha implementat una funció d'imputació personalitzada que ha permès omplir els buits de manera informada, recolzant-se en la correlació entre característiques i en el coneixement expert del domini.

Preprocessament i Normalització

El preprocessament de les dades ha estat una etapa fonamental, incloent la normalització de les característiques numèriques i la codificació de les variables categòriques. Això ha millorat significativament la qualitat de les dades per a la seva utilització en models de machine learning.

Selecció i Avaluació de Models

S'ha procedit amb la selecció de models basant-nos en la seva capacitat predictiva i la seva idoneïtat en relació amb les característiques del problema. El model SVM ha estat escollit per la seva eficiència, però l'anàlisi ha demostrat que altres models, com ara Random Forest i EBM, han ofert un millor rendiment en alguns aspectes.

Consideracions Ètiques

S'ha posat un èmfasi especial en les consideracions ètiques de l'ús d'aquests models en l'àmbit mèdic. S'ha conclòs que, tot i que els models poden oferir una orientació valuosa, no poden substituir l'avaluació mèdica professional.

Conclusions Finals

En conclusió, aquest projecte ha proporcionat una comprensió profunda dels desafiaments i oportunitats en la predicció de l'estat de salut dels pacients amb cirrosi. Ha quedat clar que, malgrat els avanços en machine learning, la complexitat dels problemes mèdics exigeix una combinació de models predictius avançats, coneixement expert del domini i una consideració cuidadosa de les implicacions clíniques de les prediccions dels models.

References

[1] Cirrhosis - symptoms and causes, 2023.

[2] Cirrhosis - symptoms, 2023.

[3] Cirrhosis patient survival prediction dataset, 2023.

[4] Liver function tests, 2023.

[5] Liver function tests - prothrombin time, 2023.

[6] Dataset information - cirrhosis patient survival, 2023.

[7] Bilirubin blood test information. Accessed: 2023-12-25.

[8] Normal bilirubin levels vs. bilirubin levels in liver cirrhosis. Accessed: 2023-12-25.

[9] What other liver diseases can cause high bilirubin? Accessed: 2023-12-25.

[10] How does cirrhosis affect bilirubin levels? Accessed: 2023-12-25.

[11] Cholesterol, 2023.

[12] Cholesterol levels: What numbers should you aim for?, 2023.

[13] The effects of cirrhosis on cholesterol levels, 2023.

[14] Cirrhosis and its impact on cholesterol, 2023.

[15] Copper in diet, 2023.

[16] Copper, 2023.

[17] Alkaline phosphatase test, 2023.

[18] Alkaline phosphatase, 2023.

[19] Prothrombin time (pt), 2023.