



UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH

Universitat Politècnica de Catalunya

FACULTAT D'INFORMÀTICA DE BARCELONA

PREDICCIÓ DE SUPERVIVÈNCIA EN  
PACIENTS AMB CIRROSI HEPÀTIC

*Grau en intel·ligència artificial*

Introducció a l'Aprendentatge Automàtic

Núria Llopart

28/12/2023

# Índex

<b>1</b>	<b>Introducció</b>	<b>3</b>
<b>2</b>	<b>Anàlisis i preprocessat de dades</b>	<b>4</b>
2.1	Anàlisi variables numèriques . . . . .	5
2.2	Anàlisi variables categòriques . . . . .	13
2.3	Recodificació de variables . . . . .	18
2.4	Tractament d'outliers . . . . .	21
2.5	Partitionat de les dades . . . . .	27
2.6	Tractament de missings . . . . .	27
2.7	Tractament del desbalanceig de classes . . . . .	31
<b>3</b>	<b>Preparació de variables</b>	<b>36</b>
3.1	Normalització . . . . .	36
3.2	Anàlisi de correlacions . . . . .	38
3.3	Anàlisi variables categòriques . . . . .	46
3.4	Estudi dimensionalitat amb PCA . . . . .	49
3.5	Variables sorolloses . . . . .	52
<b>4</b>	<b>Resum del preprocessament</b>	<b>53</b>
<b>5</b>	<b>Definició de models</b>	<b>54</b>
5.1	KNN: . . . . .	54
5.2	Arbre de decisió: . . . . .	56
5.3	SVM: . . . . .	59
5.4	EBM . . . . .	61
<b>6</b>	<b>Selecció de models</b>	<b>64</b>
6.1	KNN final . . . . .	64
6.2	Arbre de decisió final . . . . .	66
6.3	SVM final . . . . .	70
6.4	EBM final . . . . .	73
<b>7</b>	<b>Model Card: Predictor de supervivència en pacients amb cirrosi hepàtic</b>	<b>85</b>
<b>8</b>	<b>KMeans i Hierarchical Clustering</b>	<b>96</b>



# 1 Introducció

La cirrosi hepàtic és una malaltia del fetge que resulta de la cicatrització progressiva del teixit hepàtic.

En aquest treball, intentaré desenvolupar un model predictiu que utilitzi dades clíniques per predir la supervivència dels pacients amb cirrosis hepàtic. Per fer-ho, utilitzaré un conjunt de dades amb 20 característiques, sobre l'estat del pacient. La variable objectiu del model serà l'estat 'Status' que indica si el pacient ha sobreviscut o no.

El desenvolupament del projecte s'estructura en diverses fases, la primera que consta de l'exploració de les dades. A la segona fase, es farà la preparació de les dades. Seguidament, es crearan els diferents models i s'entrenaran, dels quals se'n triarà un, el que millor treballi per ser el model final. Per triar el model, es durà a terme una extensa evaluació del model i una detallada interpretació dels resultats.

Un cop el model final ja sigui decidit, aquest intentarà predir les dades del test per així poder analitzar quina exactitud té.

Un cop realitzada l'anàlisi dels resultat del test, es crearà el model card del model i finalment, es realitzarà un KMeans i un Hierarchical clustering per poder determinar si el conjunt de dades es pot dividir en grups segons la variable objectiu.

Així doncs, l'objectiu d'aquest treball no només és crear un model predictor per saber si el pacient morirà, sobreviurà o sobreviurà amb trasplantament, sinó que té com a objectiu aprendre a generar i a avaluar models i a prendre decisions i extreure conclusions al respecte.

## 2 Anàlisis i preprocessat de dades

El primer pas per a realitzar una anàlisi i un preprocessament de dades efectiu és adquirir una comprensió profunda de l'estructura del conjunt de dades. En aquest escenari particular, es disposa d'un total de 418 instàncies, cadascuna caracteritzada per 20 atributs únics.

Seguidament, es va dedicar un temps a estudiar i interpretar la informació que cada atribut conté, així com comprendre el format amb què aquesta informació es presenta.

La figura 1 ofereix un resum exhaustiu per a cada atribut present en el conjunt de dades. Aquest resum conté diversos elements clau, incloent-hi el nom de la columna, una descripció detallada de l'atribut, el seu tipus (ja sigui numèric o categòric) i, en cas de ser numèric, la corresponent unitat de mesura. Si l'atribut és de naturalesa categòrica, es proporcionen les diverses modalitats que conté.

Atribut	Descripció	Tipus	Unitats de mesura/modalitat
ID	Identificador únic	Integer	-
N_Days	Dies entre el registre i la data més recent de mort, trasplantament o temps d'anàlisi de l'estudi a juliol de 1986.	Integer	Dies
Status	Estat del pacient	Categòrica	C (censored) / CL (censored due to liver tx) / D (death)
Drug	Tipus de droga D-penicillamine o placebo	Categòrica	D-penicillamine / Placebo
Age	Edat	Integer	Dies
Sex	Sexe	Categòrica	M (male) / F (female)
Ascites	Presència de ascites	Categòrica	Y (yes) / N (no)
Hepatomegaly	Presència d'hepatomegaly	Categòrica	Y (yes) / N (no)
Spiders	Presència d'spiders	Categòrica	Y (yes) / N (no)
Edema	Presència d'edema	Categòrica	N (sense edema i sense teràpia diurètica per a l'edema) / S (edema present sense diurètics, o edema resolt amb diurètics) / Y (edema malgrat la teràpia amb diurètics)
Bilirubin	Bilirubina sèrica	Continuous	mg/dl
Cholesterol	Colesterol sèric	Integer	mg/dl
Albumin	Albumina	Continuous	gm/dl
Cooper	Coure en orina	Integer	µg/dia
Alk_Phos	Fosfatasa alcalina	Continous	U/litre
SGOT	SGOT	Continous	U/ml
Triglycerides	Triglicèrids	Integer	-
Platelets	Plaquetes per cúbic	Integer	ml/1000
Prothrombin	Temps de protrombina	Continous	segons
Stage	Estadi histològic de la malaltia	Categòrica	1/2/3/4

Figura 1: Resum atributs

Després de realitzar un resum general del conjunt de dades, es va procedir amb l'anàlisi detallada de cada variable. Aquesta fase inicial va implicar la classificació de les variables en dues categories fonamentals: les variables categòriques i les variables numèriques. En total, la base de dades es compon de 8 variables numèriques i 12 variables categòriques.

## 2.1 Anàlisi variables numèriques

A la figura 2, es presenta una visió detallada de les característiques de les variables numèriques. La taula proporciona una anàlisi completa, incloent-hi el nom de cada variable, la seva freqüència d'aparició en les instàncies, la mitjana, la desviació estàndard, el valor mínim, el valor dels tres rangs interquartils (25%, 50% i 75%) i el valor màxim.

Atribut	Count	Mitjana	std	min	25%	50%	75%	max
ID	418	209.5	120.8105	1	105.25	209.5	313.75	418
N_Days	418	1917.7823	1104.673	41	1092.75	1730	2613.5	4795
Age	418	18533.3517	3815.8451	9598	15644.5	18628	21272.5	28650
Bilirubin	418	3.2208	4.4075	0.3	0.8	1.4	3.4	28
Cholesterol	284	369.5106	231.9445	120	249.5	309.5	400	1775
Albumin	418	3.4974	0.425	1.96	3.2425	3.53	3.77	4.64
Copper	310	97.6484	85.6139	4	41.25	73	123	588
Alk_Phosphat	312	1982.6558	2140.3888	289	871	1259	1980	13862.4
SGOT	312	122.5563	56.6995	26.35	80.6	114.7	151.9	457.25
Tryglicerides	282	124.7021	65.1486	33	84.25	198	151	598
Platelets	407	257.0246	98.3256	62	188.5	251	318	721
Prothrombin	416	10.7317	1.022	9	10	10.6	11.1	18

Figura 2: Resum variaables numèriques

Seguidament, es va realitzar una anàlisi de les distribucions de cada variable.

### ID:

Atès que aquesta variable serveix com a identificador únic per a cada individu, no és necessari realitzar una anàlisi detallada de la variable, ja que cada valor és únic i no aporta informació addicional significativa.

### N\_Days:

La variable N\_Days indica els dies entre el registre i la data més recent de mort, trasplantament o temps d'analisi de l'estudi a juliol de 1986. La distribució de la variable es mostra a la figura 3.

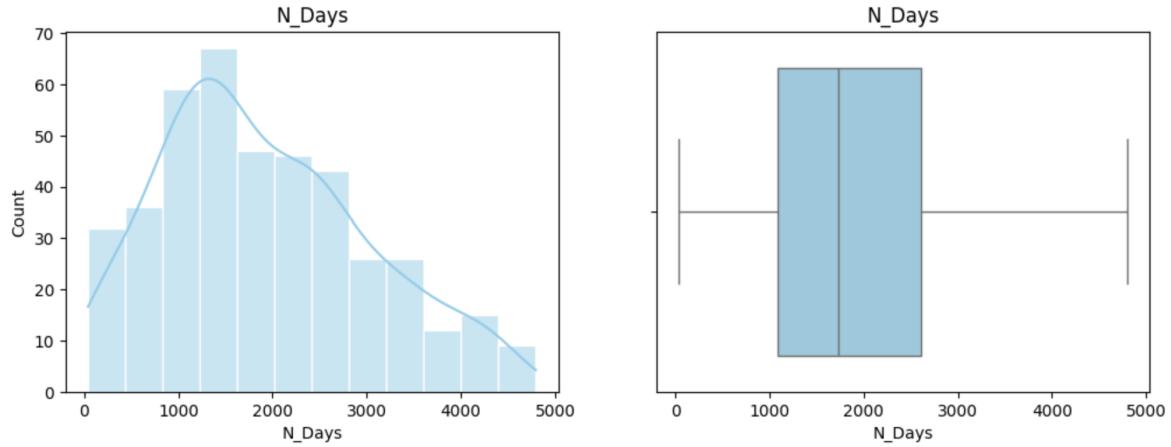


Figura 3: Distribució variable N\_Days

En aquest cas, els valors de la variable oscil·len entre 41 a 4795 dies. La mitjana d'aquests valors se situa en 1917.7823 dies, aproximadament equivalent a uns 5 anys. L'anàlisi de l'histograma revela que la major concentració de valors es troba en el rang de 0 a 2000 dies, amb un pic notable entre els 1000 i 2000 dies.

En referència al boxplot, aquest també ressalta aquesta tendència cap als valors més petits. Es pot notar que la línia representant la mitjana es troba més a prop del primer quartil, reforçant la idea que les dades estan esbiaixades cap a valors més baixos. A més a més, tal com es pot apreciar al boxplot, a simple vista aquesta variable no conté valors atípics o outliers.

#### Age:

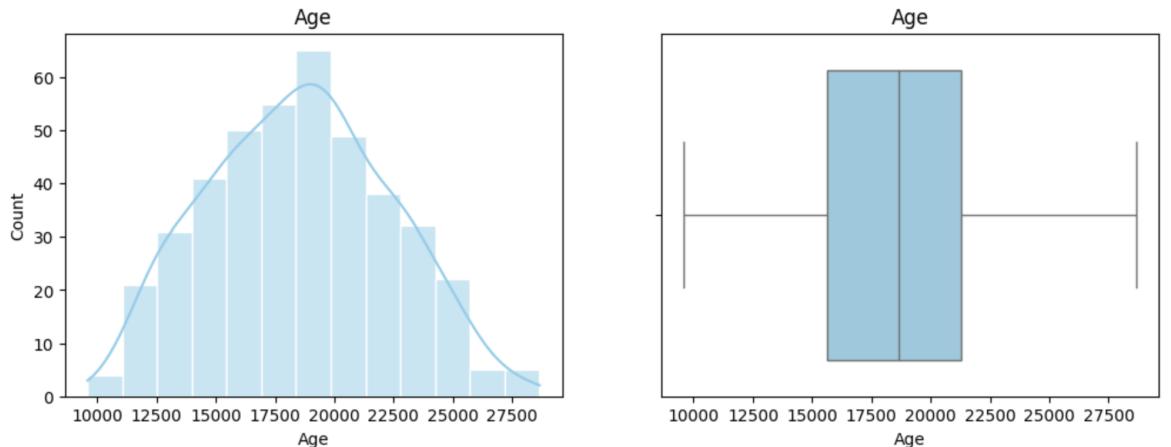


Figura 4: Distribució variable age

La variable age ens indica l'edat de les persones en dies.

En aquesta anàlisi, es fa evident que la distribució presenta similituds amb una distribució gaussiana, malgrat no ser perfectament simètrica. No obstant, podem detectar un pic entre els 17500 i els 20000

dies (equivalent a les edats entre els 48 i els 55 anys). En aquest cas, la mitjana de la distribució se situa en 18533.3571 dies (equivalent a 51 anys), amb els valors que oscil·len entre 9598 dies (equivalent a 26 anys) i 28659 dies (equivalent a 78 anys). Es pot observar com la variable no conté outliers.

### Bilirubin:

La variable bilirubin indica la bilirubina sèrica del pacient en mg/dl.

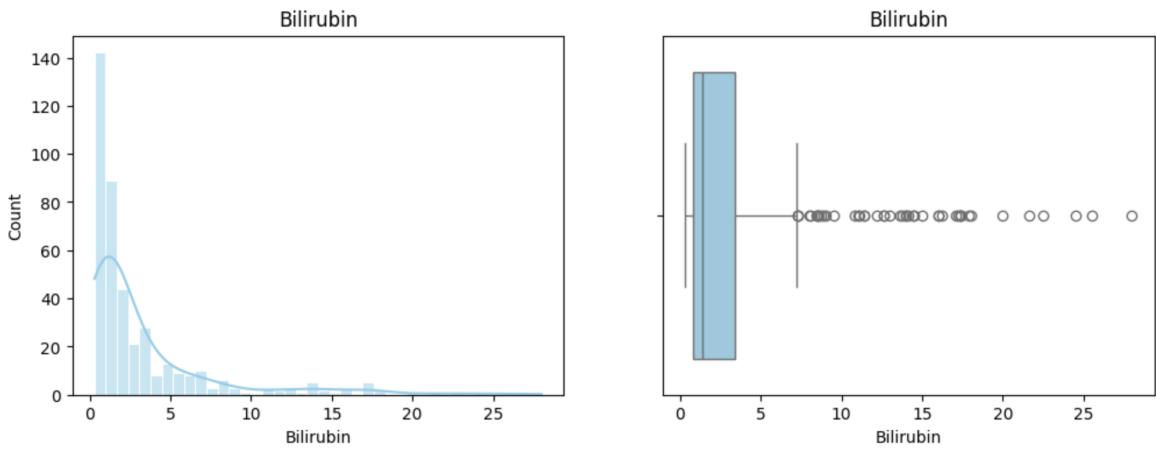


Figura 5: Distribució variable Bilirubin

A l'histograma de la figura 5 es pot veure una distribució logarítmica on el pic es troba entre els valors 0-2. En aquesta variable, la mitjana es troba en 3.2208 i els seu rang se situa entre 0.3 i 28.

Amb el boxplot, es fa evident la presència d'un nombre significatiu d'outliers, tots els valors atípics són majors al tercer rang interquartil. Més endavant es realitzarà una anàlisi exhaustiva per poder determinar com tractar aquests valors.

### Cholesterol

La variable cholesterol ens indica el nivell de colesterol del pacient.

La distribució presenta un pic concentrat en els valors compresos entre 250 i 450. No obstant això, després d'aquest punt, l'ocurrència de valors disminueix de manera notable. La mitjana de la distribució es troba en 369.51 amb una desviació estàndard de 231.94. Ens valors varien en un rang que s'estén des de 120 a 1775.

Observant el boxplot es pot apreciar com els rang interquartils se situen a prop del pic. A més a més, es pot observar una notable presència d'outliers que se situen més enllà del tercer quartil, més profunda sobre la dispersió i la presència de valors extrems en la variable en consideració.

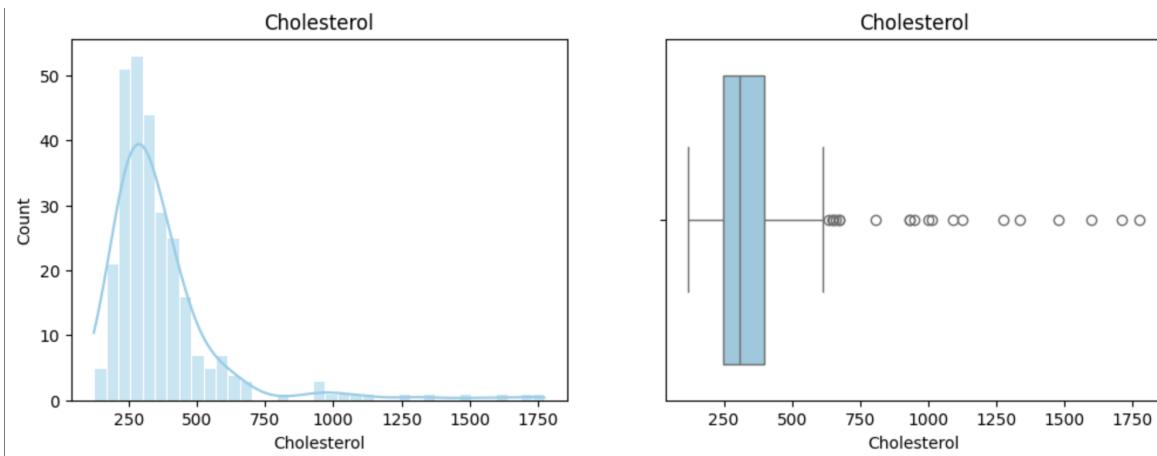


Figura 6: Distribució variable cholesterol

### Albumin:

La variable Albumin indica el nivell d'Albumina en gm/dl.

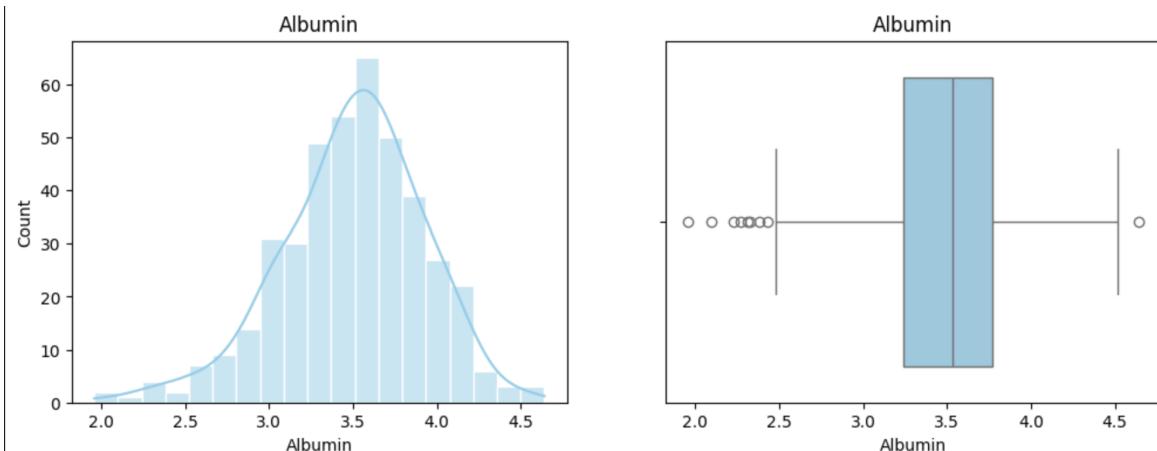


Figura 7: Distribució variable Albumin

La variable Albumin presenta una distribució que recorda a una campana de Gauss. La mitjana es troba en 3.4974, mentre que l'abast de la variable va des de 1.96 fins a 4.64.

En analitzar el boxplot, es pot observar una quantitat significativa d'outliers, la majoria es troben per sota del primer quartil, i un únic valor supera el tercer quartil.

### Copper:

La variable Copper ens indica el coure en orina del pacient en  $\mu\text{g}/\text{dia}$ .

Analitzant la distribució podem detectar un pic entre els valors 4 i 100. A partir d'aquest pic, la

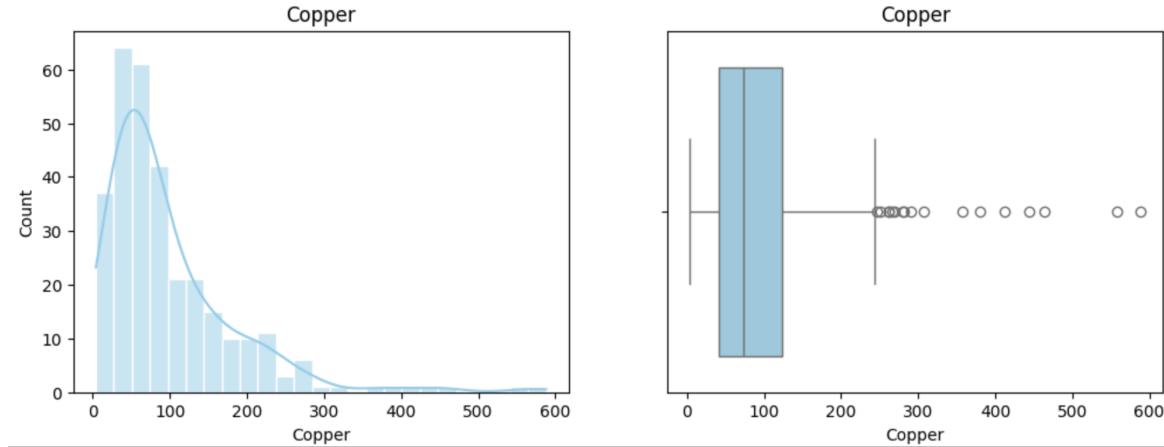


Figura 8: Distribució variable Cooper

gràfica té un comportament logarítmic. En aquesta variable trobem un rang de valors considerablement ampli, variant des d'un mínim de 4 fins a un màxim de 588. Tot i això, la mitjana se situa en 97.6484 amb una desviació estàndard de 85.61.

Observant el boxplot podem observar com aquest presenta una gran quantitat d'outliers en els valors superior al tercer quartil.

### Alk\_Phosphat:

La variable ens indica la quantitat de fosfatasa alcalina en U/litre.

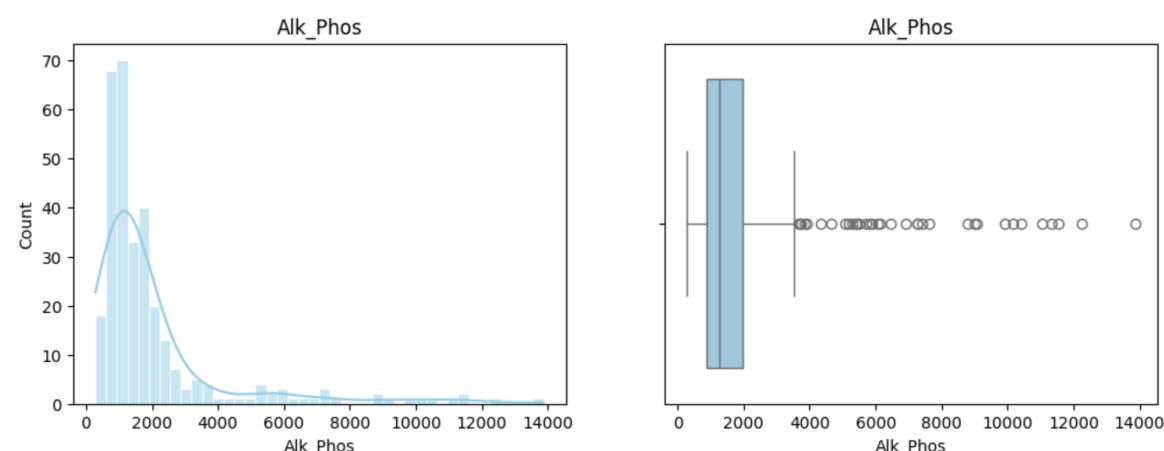


Figura 9: Distribució variable Alk\_Phosphat

En aquest cas, es pot observar un pic destacat en els valors inicials, concretament entre 0 i 2000. Després d'aquest punt, la distribució mostra una disminució. Malgrat això, el rang de valors abasta des de 289 fins a 13862.4. La mitjana se situa en 1982.6558.

Observant el boxplot podem apreciar una gran quantitat d'outliers coincident amb els valors més elevats.

### **SGOT:**

La variable SGOT ens indica la quantitat d'SGOT del pacient per U/ml.

Els primers valors, fins a 150 aproximadament de la variable tenen una distribució similar a la

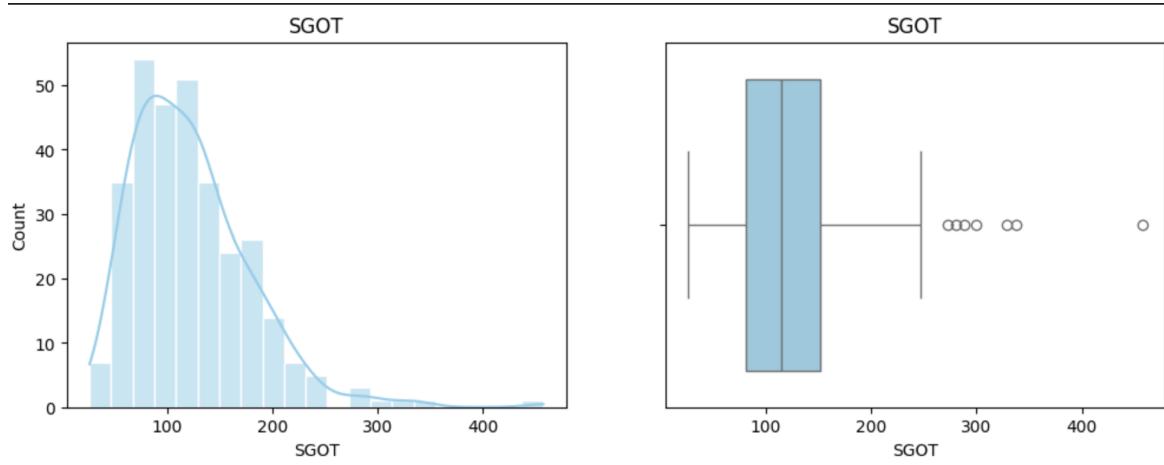


Figura 10: Distribució variable SGOT

gaussiana però a mesura que aquests valors augmenten, disminueix la seva freqüència. La mitjana se situa en 122.56 amb una distribució estàndard de 56.7. A més a més, el rang de valors s'esten de 26.35 fins a 457.25.

Observant el boxplot podem detectar uns quants valors atípics situats per sobre el tercer quartil.

### **Triglycerides:**

Aquesta variable idica el triglicèrids que té el pacient.

Observant la distribució de la variable es pot relacionar la distribució dels valors inicials (entre 33 i uns 200) similar a una distribució normal. Després del pic, aquests valors disminueixen significativament. La mitjana se situa al 124.7 amb una desviació estàndard de 65.15. El rang de valors és extens, amb un mínim de 33 i un màxim de 598.

Analitzant el boxplot, es pot apreciar l'efecte del pic ja que els quartils se situen cap als primers valors. A més a més, podem detectar uns quants outliers superiors al tercer quartil.

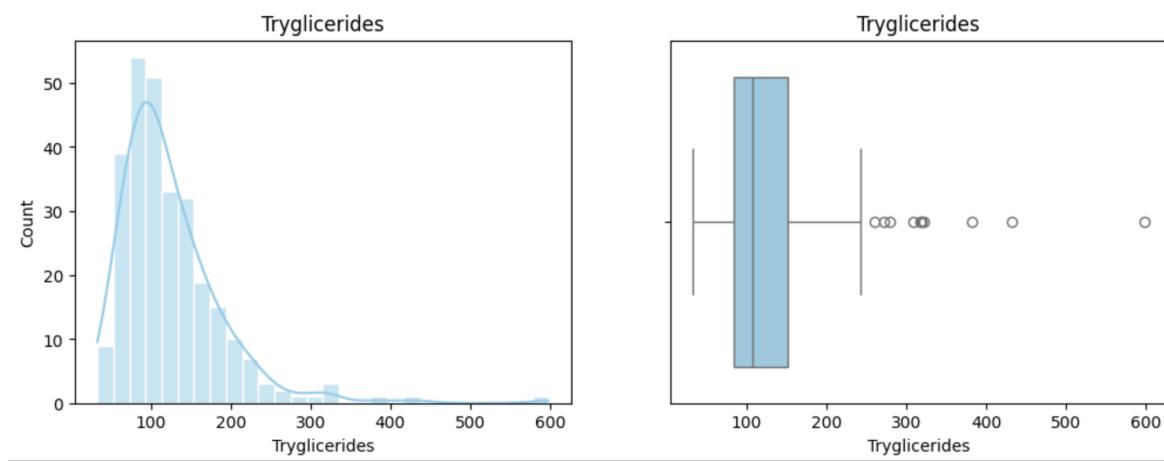


Figura 11: Distribució variable Tryglicerides

#### Platelets:

La variable Platelets indica la quantitat de plaquetes per cúbic que té el pacient amb unitats de ml/1000.

Analitzant l'histograma es pot detectar una distribució que sembla ser lleugerament esbiaixada cap

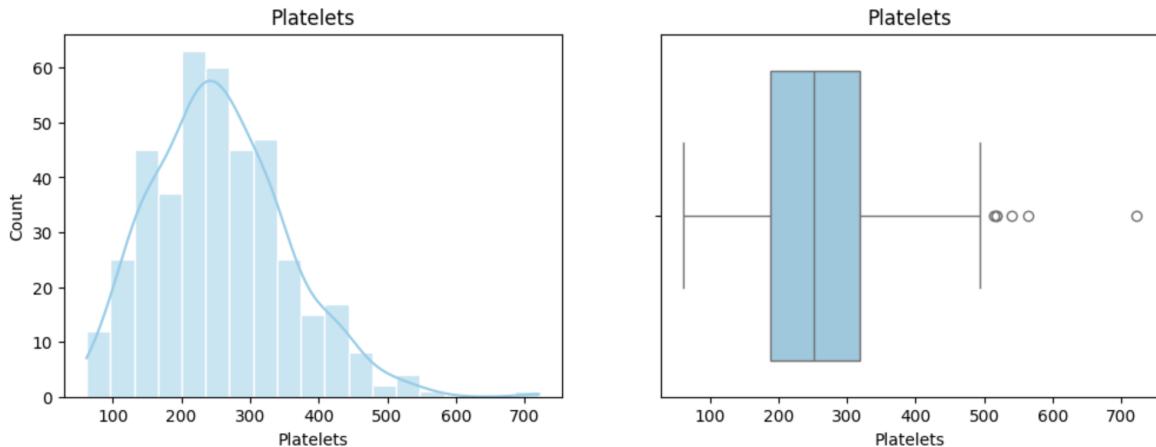


Figura 12: Distribució variable Platelets

a la dreta, indicant que hi ha més valors alts que no són tan comuns. La mitjana se situa en 257.03 i amb una desviació estàndard de 98.33. El rang dels valors se situa entre 62 i 721, el que mostra un rang de valors extens.

Observant el boxplot, podem detectar alguns outliers situats per sobre del tercer rang quartil. També es pot observar com la majoria de valors es centren entre aproximadament 150 i 400.

### **Prothrombin:**

La variable prothrombin indica el temps de protrombina en segons.

Si s'analitza la distribució es pot veure com aquesta és lleugerament esbiaixada cap a la dreta però

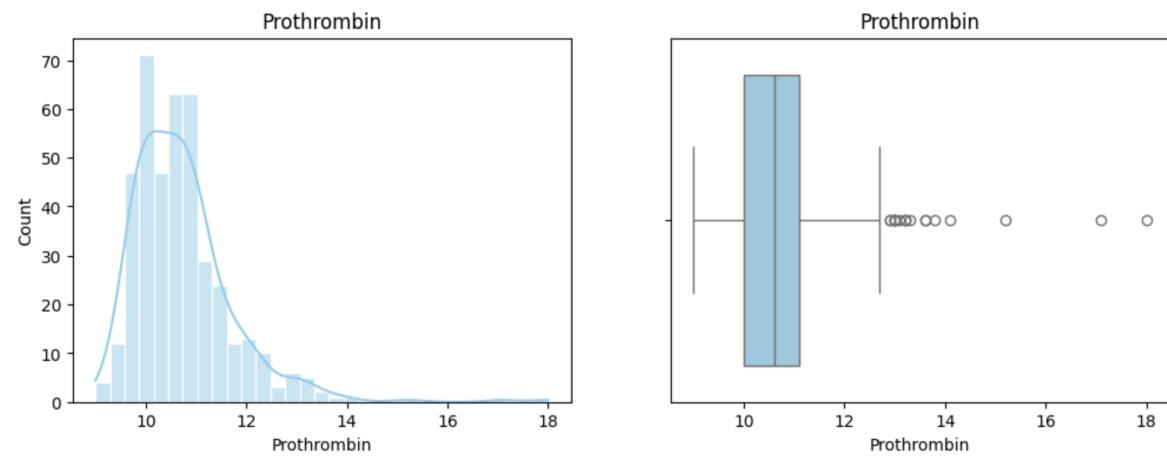


Figura 13: Distribució variable Prothrombin

de forma poc pronunciada.

En aquest cas, la mitjana se situa en 10.13 amb una desviació estàndard 1.022 amb un rang que va de 9 a 11.1.

Amb el boxplot es poden detectar uns quants outliers superiors al tercer quartil.

## 2.2 Anàlisi variables categòriques

### Status:

La variable "status" indica l'estat del pacient i és la variable objectiu. Aquesta variable té tres categories: 'C' (censored), és a dir, que el pacient ha sobreviscut; 'CL' (censored due to liver tx), indicant que la supervivència es deu a un trasplantament de fetge; i 'D' (death), és a dir, que el pacient ha mort.

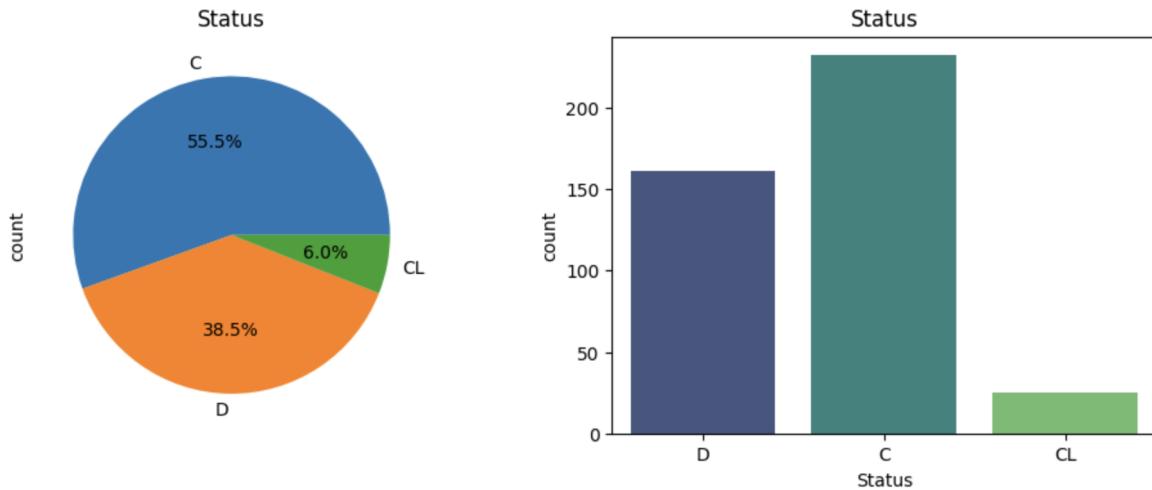


Figura 14: Variable status

Observant les gràfiques, es fa evident un desequilibri notable. La classe 'C' és predominant, abarcant més de la meitat de les observacions, mentre que la classe 'CL' és la menys freqüent amb tan sols un 6% de la representació.

### Drug:

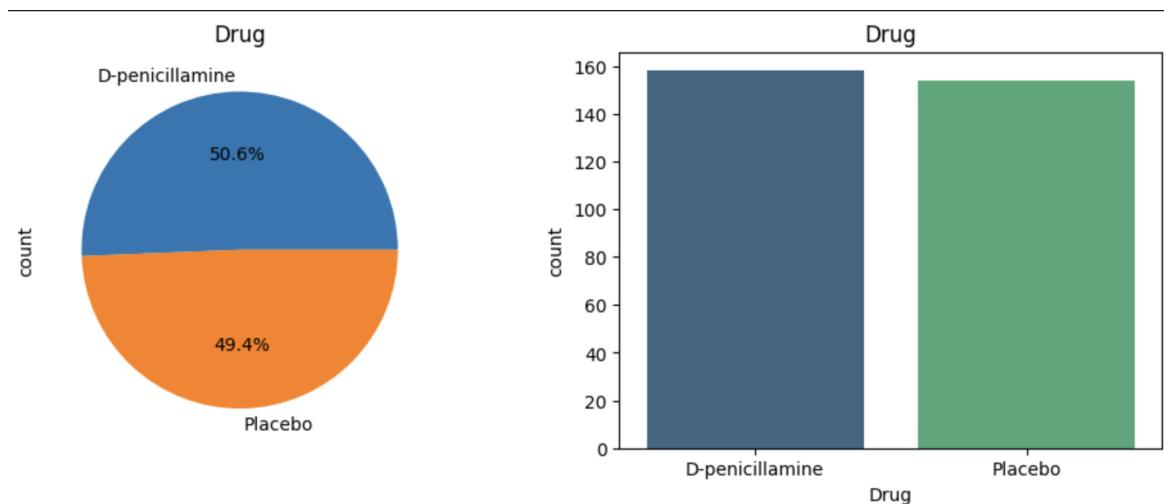


Figura 15: Variable drugs

La variable 'drug' ens indica si el pacient ha rebut D-penicillamine o placebo.

Observant les gràfiques podem determinar que les classes estan balancejades, no obstant això, podem observar la presència d'una tercera classe minoritària corresponent als valors faltants o 'missings'.

#### Sex:

La variable sex està format per dues classes, la classe 'Female' i la classe 'Male', és a dir, la classe dona i la classe home.

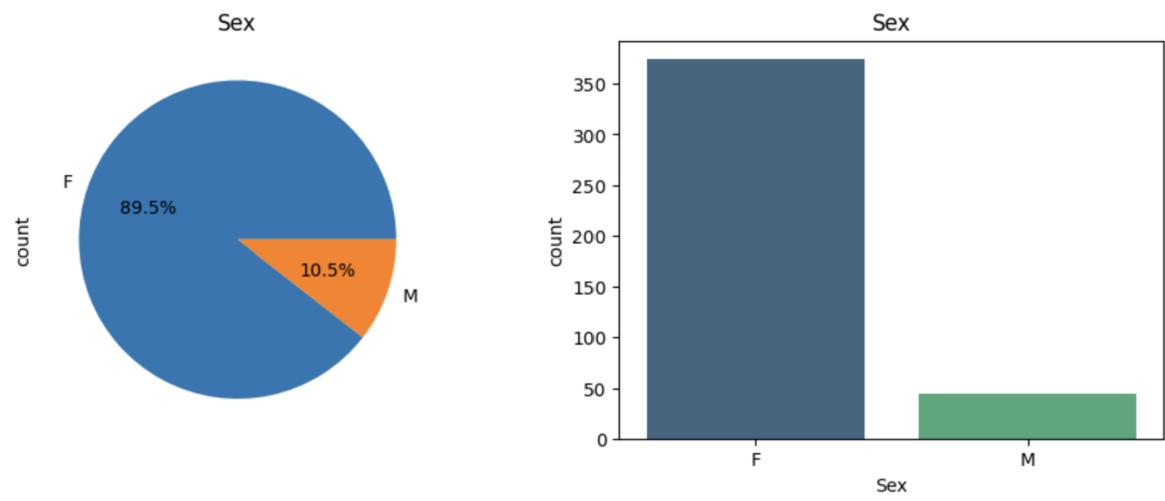


Figura 16: Variable sex

En aquest cas, es fa evident un notable desequilibri. La classe majoritària és 'Female', amb un 89.5% de la representació, en contrast amb només un 10.5% de la representació de la classe 'Male'.

#### Ascites:

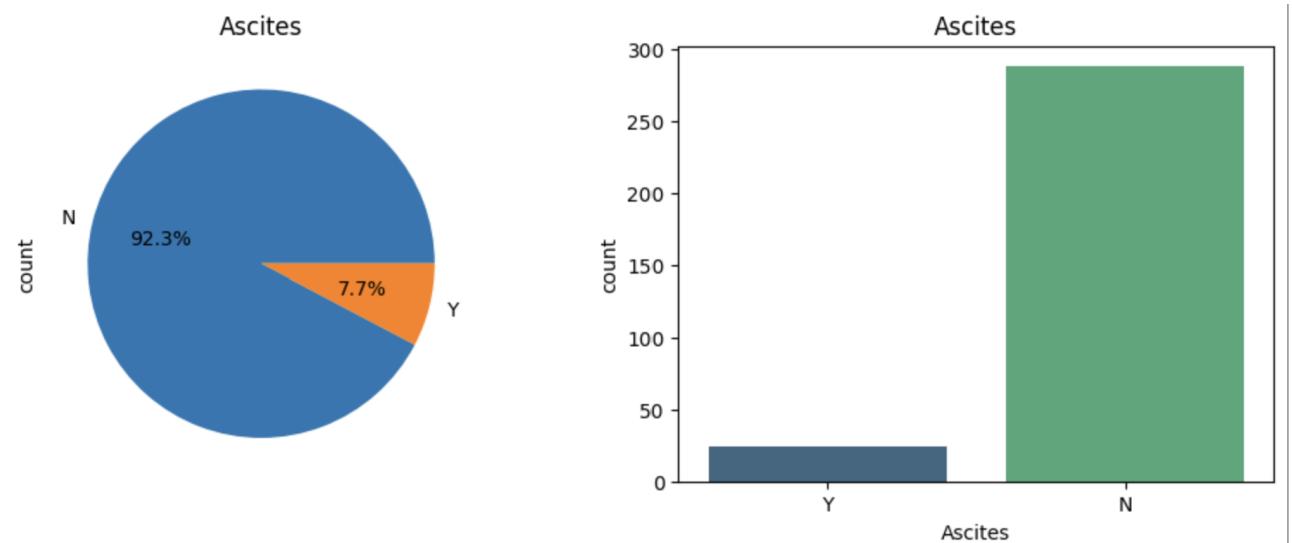


Figura 17: Variable Ascites

La variable ascites ens mostra si hi ha presència d'ascites. En aquest cas, la variable conté dues classes o 'Y', sí, o 'N', no. A la gràfica es pot veure un desbalanceig evident on la classe majoritària es la 'N', no, amb un 92.3% de la representació total.

### Hepatomegaly:

La variable Hepatomegaly determina si el pacient té presència d'hepatomegaly al cos. Les dues classes de la variable són sí o no (Y/N).

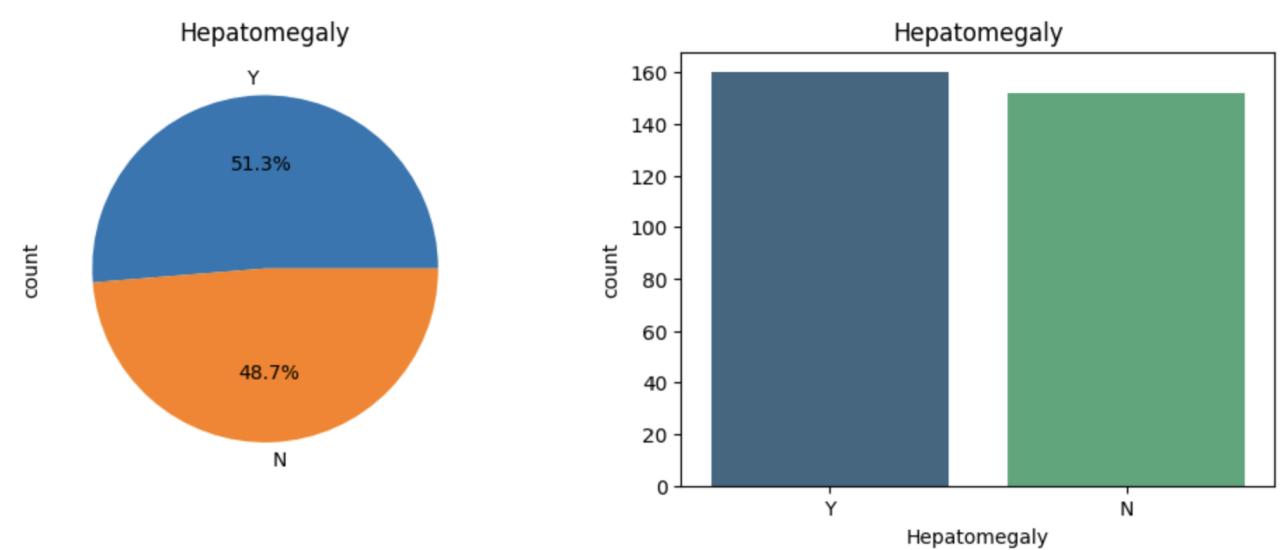


Figura 18: Variable Hepatomegaly

Observant la figura es pot determinar que la variable està balancejada malgrat que la classe 'Y' contingui el 51.3% dels valors. No obstant això, aquesta diferència es minoritària i irrelevante.

### Spiders:

Aquesta variable indica, amb dues classes, si hi el pacient té spiders o no. Les dues classes representatives són la classe 'Y' i la classe 'N' (Sí/No).

Observant les gràfiques es fa evident un desbalanceig tot i que aquest no és tan notable com en casos anteriors. La classe majoritària és la classe No amb un 71.2% de la representació total.

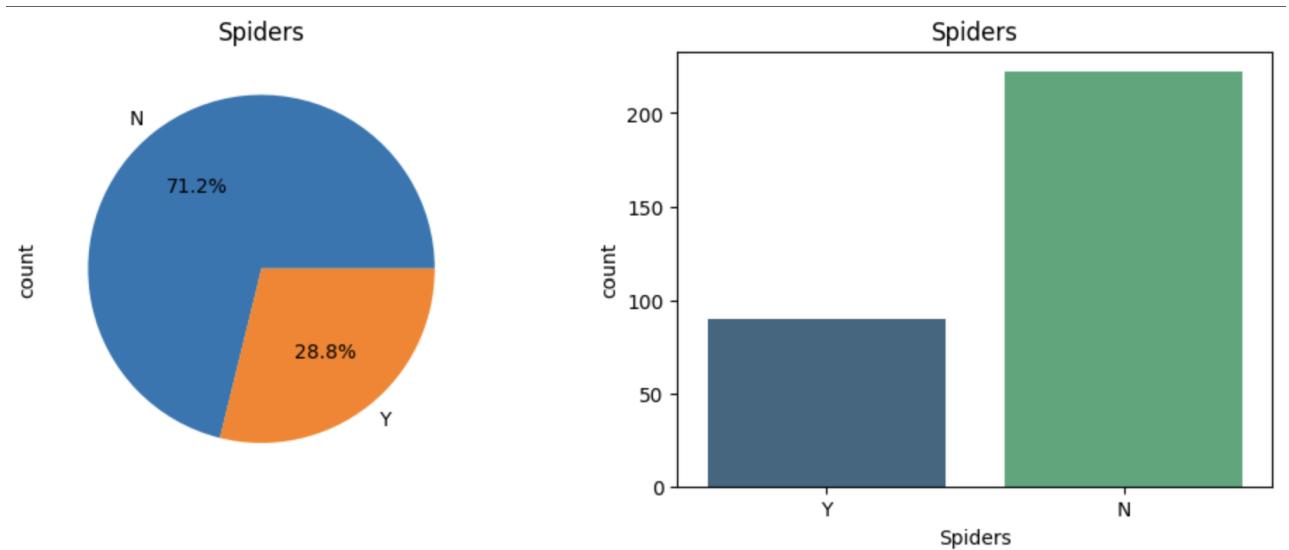


Figura 19: Variable Spiders

#### **Edema:**

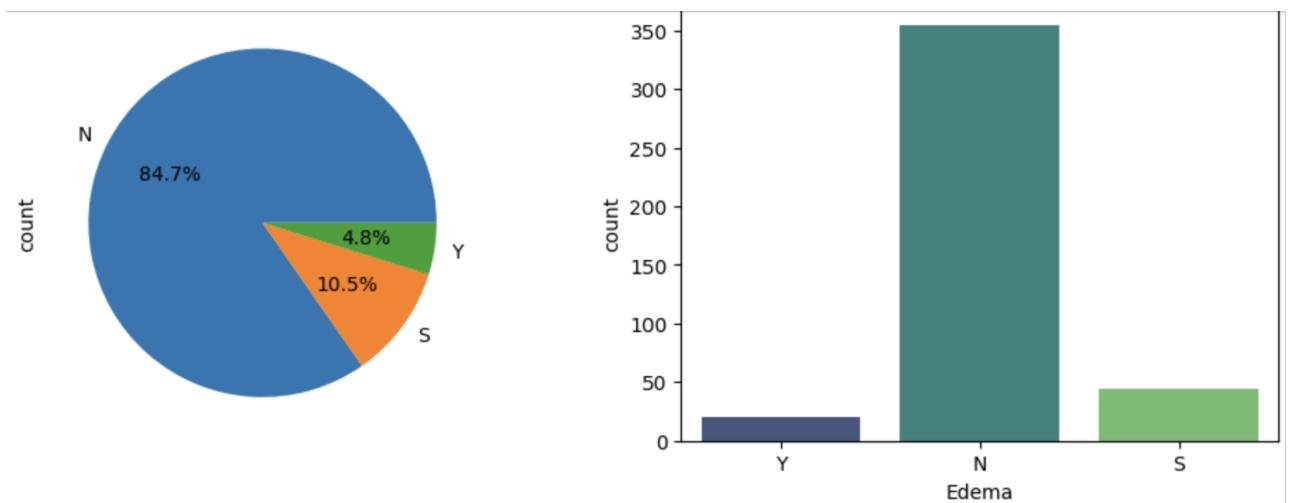


Figura 20: Variable Edema

La variable Edema ens indica si hi ha presència d'edema al cos. En aquest cas, tenim tres classes: N representant sense edema i sense teràpia diürètica per a l'edema, S representant la classe on edema està present sense diüètics o edema resolt amb diüètics i finalment la classe Y que indica que el pacient té edema malgrat la teràpia amb diüretics.

Analitzant la figura 20, es pot notar una classe majoritària que conté el 84.7% de tots els valors. Aquesta classe és la classe 'N'. Les altres dues classes són minoritàries amb un percentatge del 10.5% de la classe S i un 4.8% per la classe Y.

### **Stage:**

La variable stage ens indica l'estudi histològic de la malaltia.

En aquest cas la variable està representada per quatre nombres 1, 2, 3, 4. Aquests nombres contenen un ordre, no especificat en la descripció de la base de dades.

Analitzant la figura 21 es pot notar una classe minoritària, la classe 1, amb un 5.1% dels va-

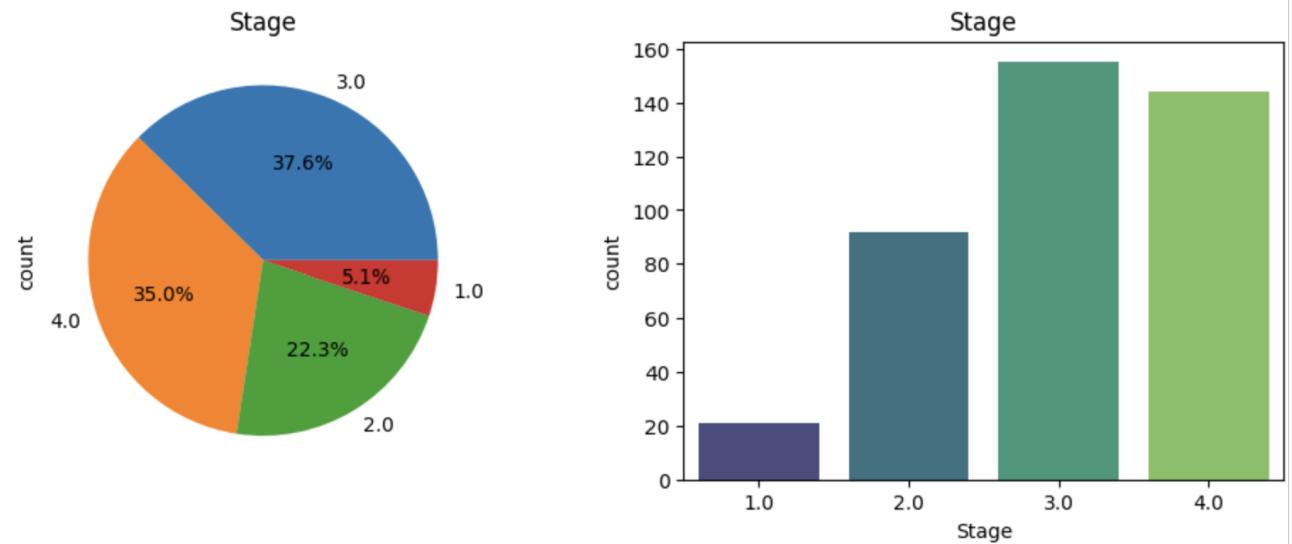


Figura 21: Variable stage

lors i dues classes majoritàries, les classes 3 i 4 amb un 37.6% i 35% dels valors respectivament. Finalment, hi ha la classe 2 que es troba en un punt entremig amb un 22.3% dels valors. Per tant, es pot observar un desbalanceig en les classes.

## 2.3 Recodificació de variables

Examinant les dades, vaig observar que algunes variables presentaven unitats poc freqüents, especialment les variables 'age' i 'N\_Days', les quals expressaven les seves mesures en dies, tot i que moltes vegades els seus valors eren significativament alts.

Amb l'objectiu de simplificar l'anàlisi i millorar la comprensió del conjunt de dades, vaig prendre la decisió de convertir les unitats d'aquestes variables a anys. En particular, vaig denominar la nova variable que reflecteix la durada en anys com a 'N\_years', reemplaçant així la variable original 'N\_Days'. D'aquesta manera, vaig millorar la interpretabilitat de les dades i vaig proporcionar una perspectiva temporal més intuïtiva.

Un cop recodificades les variables, vaig voler observar si canviar les unitats havia generat algun canvi en la distribució.

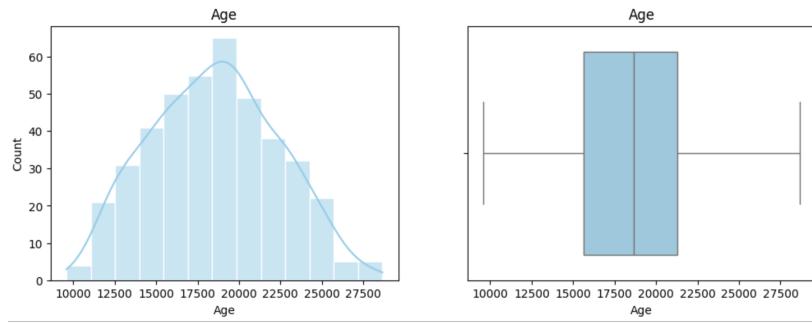


Figura 22: Distribució variable age en dies

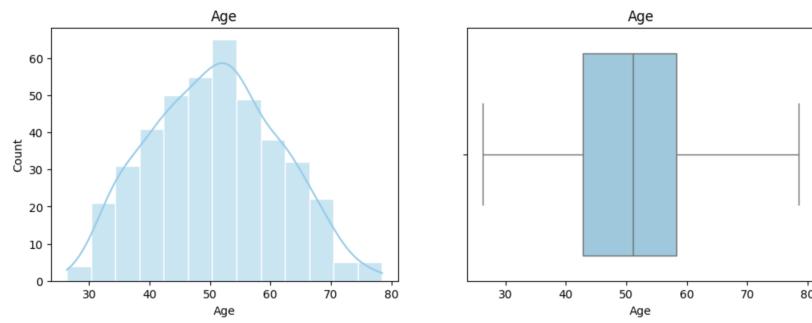


Figura 23: Distribució variable age en anys

Tal com era d'esperar, en cap cas la distribució ha canviat, simplement els nombres de l'eix de les X han disminuït fent que siguin més interpretables.

Abans d'aquest canvi, ens costava determinar el rang de les variables a simple vista ja que al tractar-se d'una quantitat tan elevada de dies, es fa estrany no parlar d'anys. Ara però, podem determinar que l'edat més elevada és 80 anys i la quantitat de dies més elevada és 12 anys.

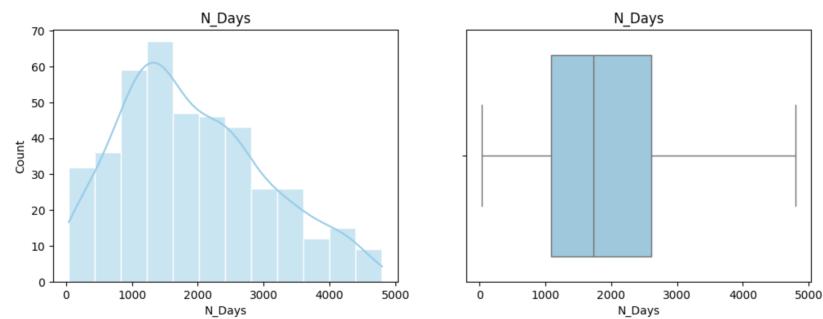


Figura 24: Distribució variable N\_Days en des

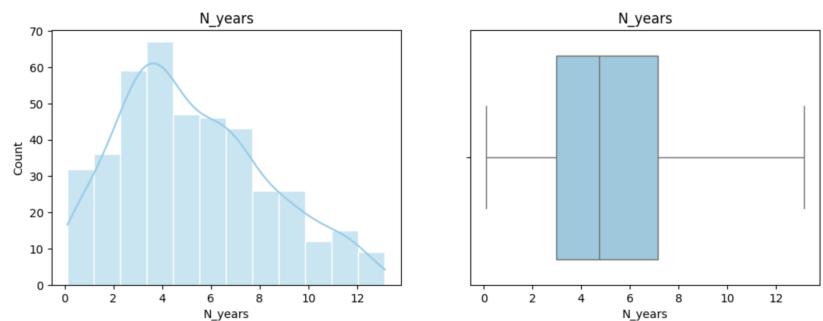


Figura 25: Distribució variable n\_years

Variable	Recodificació
Status	CL --> 1 D --> 2 C --> 3
Drug	Placebo --> 1 Dpenicillamine --> 2
Sex	F --> 1 M --> 2
Ascites	N--> 1 Y --> 2
Hepatomegaly	Y--> 1 N --> 2
Edema	N--> 1 S--> 2 Y --> 3
Spiders	Y--> 1 N --> 2

Figura 26: Label Encoding

A més a més, per poder aplicar algoritmes d'imputació i de balanceig, vaig haver de passar les variables categòriques a numèriques. Per fer-ho, vaig escollir aplicar el label encoding ja que d'aquesta manera, no crea noves variables i les dades segueixen sent fàcils d'interpretar sense necessitat de mapejar abans.

Vaig decidir realitzar el meu propi label encoding per assegurar-me d'aquesta forma, que es realitzava correctament. A la imatge de continuació, es pot observar com es van recodificar les diferents categories a nombres.

## 2.4 Tractament d'outliers

Per tractar els outliers el primer pas va ser detectar en quines variables es troben. Amb els boxplots de l'anàlisi univariant hem pogut detectar-los, no obstant, vaig optar per calcular-los mitjançant el rang interquartil per a cada variable. A la figura 27 es mostren els dos rangs interquartils per a cada variable (el límit superior, corresponent al 75% i el límit inferior, corresponent al 25%), juntament amb la quantitat total d'outliers de cada variable segons aquests rangs.

En aquest cas, hauria d'analitzar cada variable individualment i considerar si els seus outliers

Atribut	Límit inferior	Límit superior	Outliers
Bilirubin	-3.1	7.3	46
Cholesterol	46.625	583.625	14
Albumin	2.6788	4.4688	13
Copper	-58.5	201.5	19
Alk_Phosphatase	-519.5	3132.5	15
SGOT	-13.175	228.625	6
Tryglicerides	2.5	206.5	9
Platelets	30.875	481.875	3
Prothrombin	8.5	12.5	12

Figura 27: Outliers variables numèriques

són realment anomalies o simplement valors que es desvien dels altres però que són valors vàlids. Aquesta tasca s'hauria de fer consultant a un expert, desafortunadament, com que no tinc accés a un expert que pugui interpretar els valors de la base de dades, vaig optar per eliminar tots els valors que no es trobaven dins del rang determinat per la tècnica IQR.

Un cop eliminats els outliers vaig tornar a analitzar la distribució de les variables comparant-les amb les distribucions inicials.

## Bilirubin

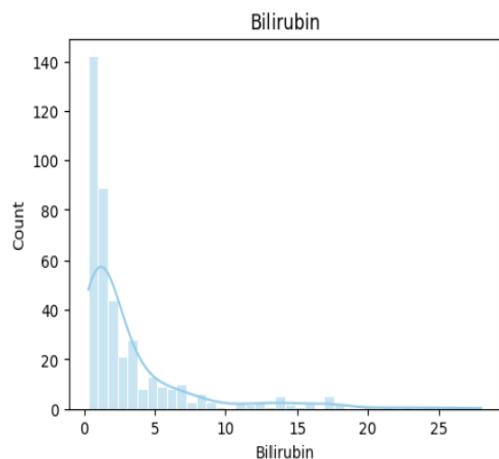


Figura 28: Distribució Bilirubin amb outliers

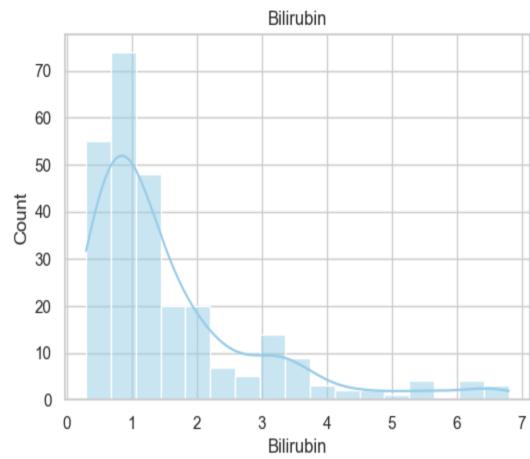


Figura 29: Distribució Bilirubin sense outliers

Figura 30: Distribucions Bilirubin

Com es pot observar, després d'eliminar els outliers, podem analitzar amb més detall la distribució de la variable. S'aprecia un pic als primers valors que disminueix a mesura que augmenta el valor de X.

## Cholesterol

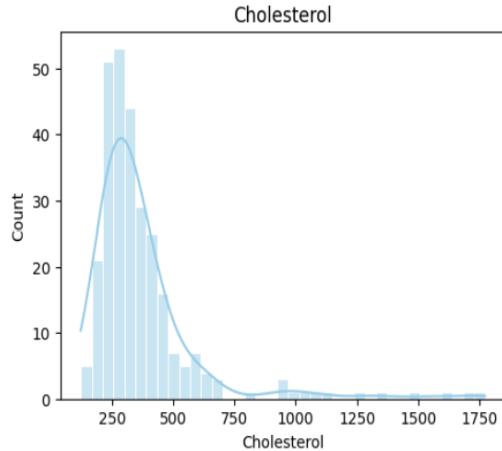


Figura 31: Distribució cholesterol amb outliers

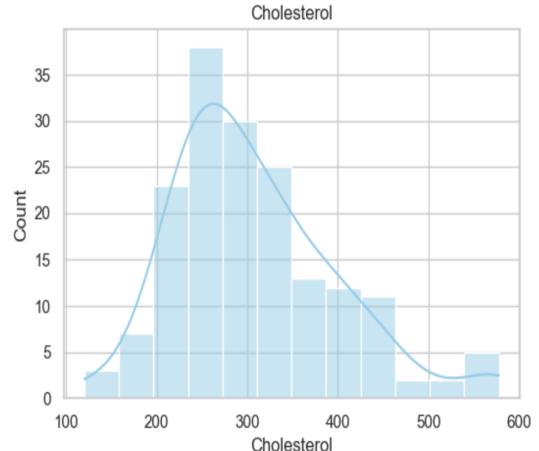


Figura 32: Distribució cholesterol sense outliers

Figura 33: Distribucions cholesterol

Un cop eliminats els outliers, podem veure com el pendent de la distribució ha disminuït, ja que els valors s'han acotat.

## Albumin

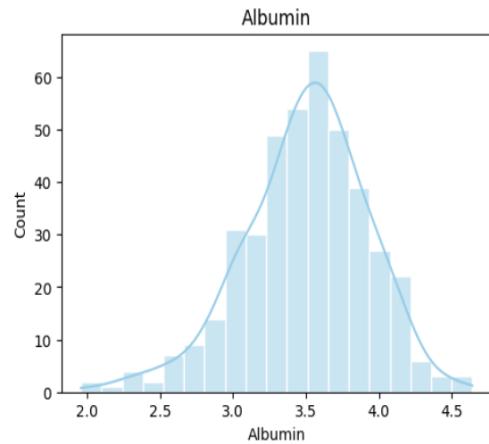


Figura 34: Distribució albumin amb outliers

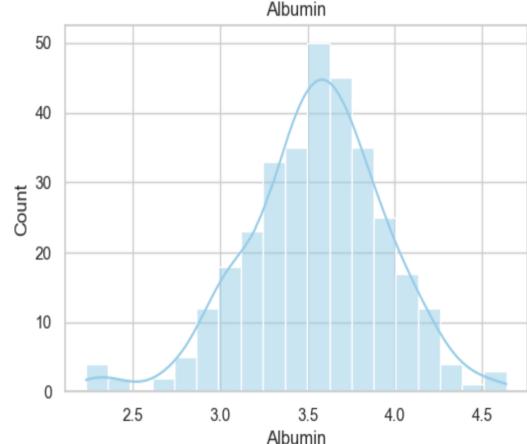


Figura 35: Distribució albumin sense outliers

Figura 36: Distribucions albumin

En aquest cas, no podem apreciar un canvi tan significatiu en la distribució de la variable ja que tan sols s'eliminen valors molt propers als límits que no alteraven la distribució inicial.

## Copper

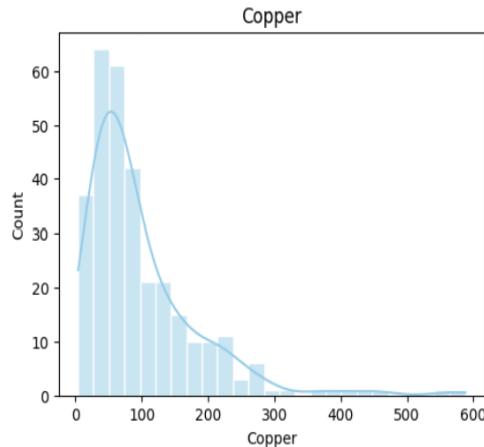


Figura 37: Distribució Copper amb outliers

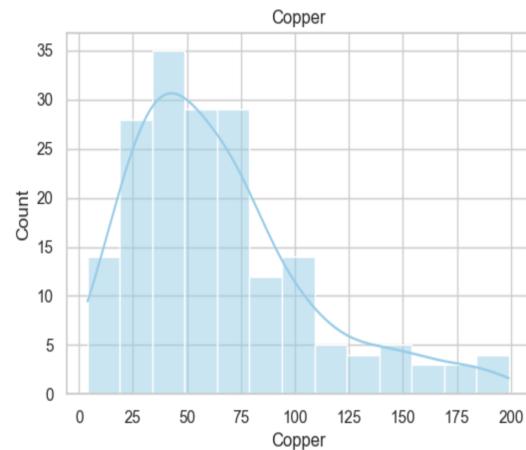


Figura 38: Distribució Copper sense outliers

Figura 39: Distribucions Copper

Un cop eliminats els outliers, podem observar com el pendent del pic disminueix ja que s'han eliminat els valors més elevats de la distribució.

## Alk\_Phosphat

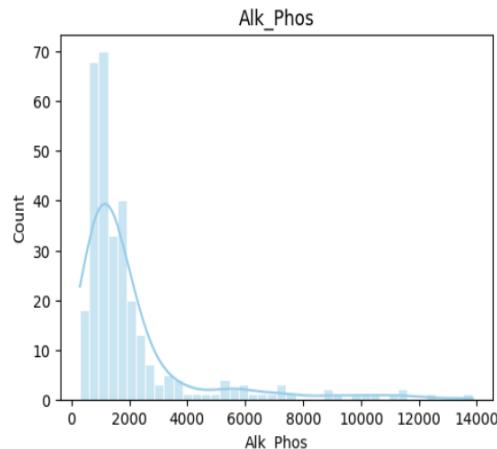


Figura 40: Distribució Alk\_Phosphat amb outliers

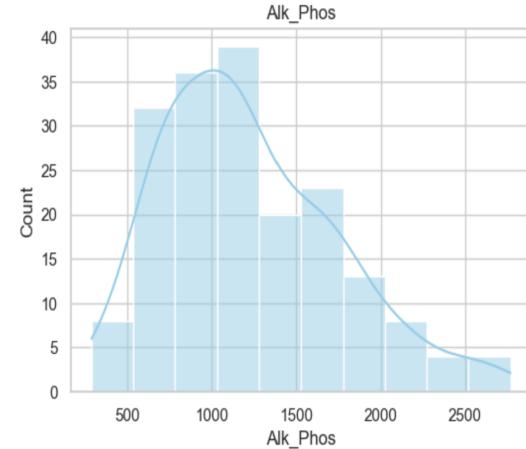


Figura 41: Distribució Alk\_Phosphat sense outliers

Figura 42: Distribucions Alk\_Phosphat

En el cas de la variable Alk\_Phosphat, si elimino els outliers, es pot apreciar com es redueix el rang de valors, quedant-me únicament amb els primers valors. D'aquesta forma, podem observar un pic esbiaixat cap a la dreta però amb la cua reduïda respecte l'altre distribució amb outliers.

## SGOT

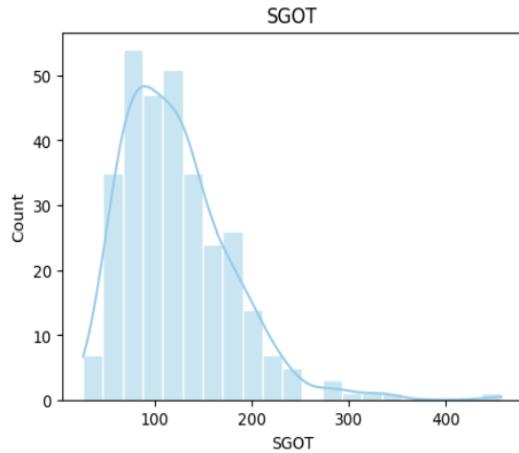


Figura 43: Distribució SGOT amb outliers

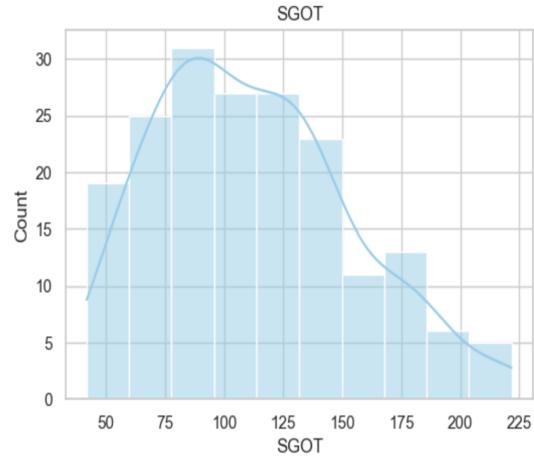


Figura 44: Distribució SGOT sense outliers

Figura 45: Distribucions SGOT

Per la variable SGOT, podem apreciar que s'han esborrat una gran quantitat de valors alts, el que fa que ara la distribució segueixi sent un pic però amb una pendent quasi desaparcebuda. No obstant això, les dades segueixen desbiaixades cap a la dreta, encara que, sense els outliers, de manera menys pronunciada.

## Tryglicerides

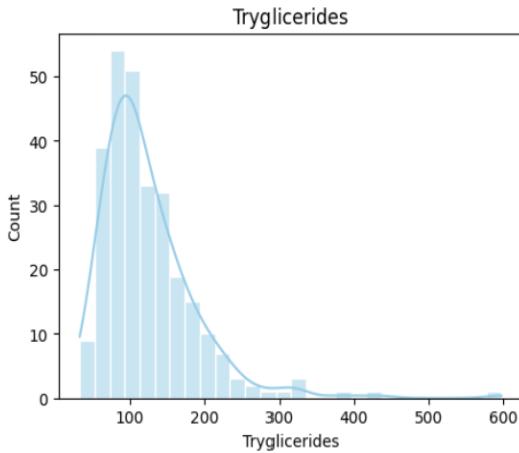


Figura 46: Distribució tryglicerides amb outliers

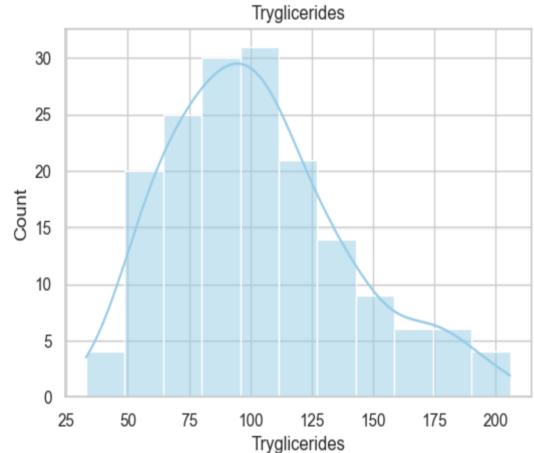


Figura 47: Distribució tryglicerides sense outliers

Figura 48: Distribucions tryglicerides

Un cop trets els outliers, es pot apreciar com la distribució segueix esbiaixada cap a la dreta però amb menys valors i de forma menys notable. Per altra banda, el pic de pujada continua tenint molta pendent.

## Platelets

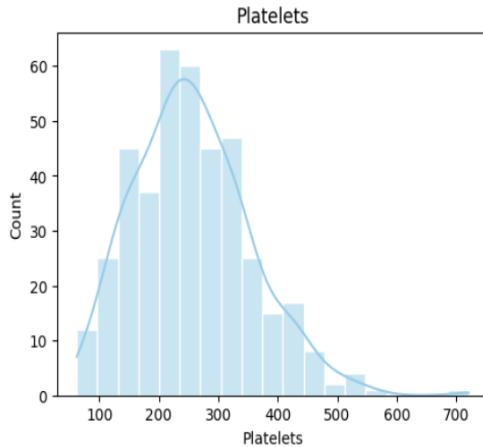


Figura 49: Distribució Platelets amb outliers

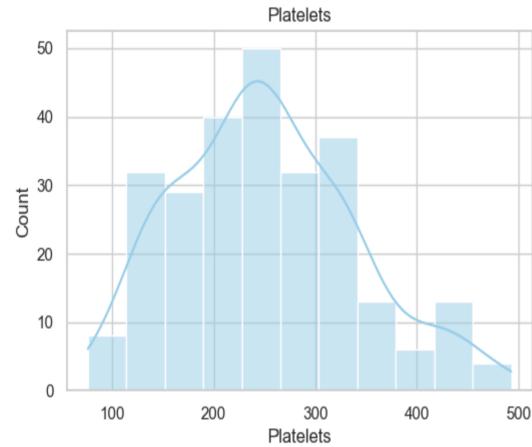


Figura 50: Distribució Platelets sense outliers

Figura 51: Distribucions Platelets

En aquest cas, el fet de treure els valors superiors, fa remarcable els pics relatius de la distribució. Per altra banda, les dades segueixen estan desbiaixades cap a la dreta però com ja ha passat en altres variables, la cua és menor.

## Prothrombin

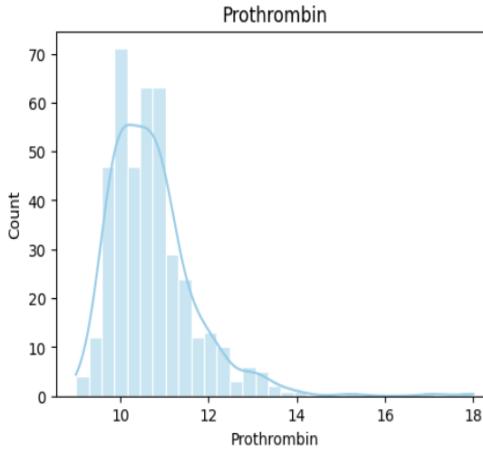


Figura 52: Distribució Prothrombin amb outliers

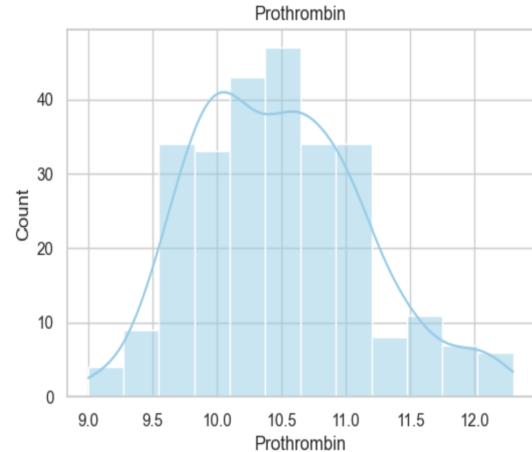


Figura 53: Distribució Prothrombin sense outliers

Figura 54: Distribucions Prothrombin

Un cop s'han eliminat els valors atípics de la variable, la distribució es presenta clarament. En aquest context, es pot observar que la variable mostra una tendència mínima a l'esbiaixament, en contrast amb la distribució de la variable amb outliers, la qual presenta un esbiaixament cap a la dreta. D'altra banda, es distingeix un pic en la distribució sense outliers, però amb una pendent lleugerament menys pronunciada.

## 2.5 Particionat de les dades

Després de gestionar els outliers, vaig prendre la decisió de dividir les dades en un conjunt d'entrenament i un conjunt de proves. En el moment de realitzar aquesta partició, vaig optar per reservar el 80% de les dades per a l'entrenament i el 20% per a les proves. Aquesta elecció es va basar en la consideració que el conjunt de dades original no disposava d'una gran quantitat de mostres, i l'afegiment d'una tercera partició per a la validació podria afectar significativament la quantitat de dades disponibles per a l'entrenament o les proves.

Vaig considerar que una distribució del 80-20 era una opció equilibrada ja que permet dedicar una gran part del conjunt de dades per l'entrenament, al mateix temps que es reserva una porció significativa de dades per poder avaluar l'eficàcia del model.

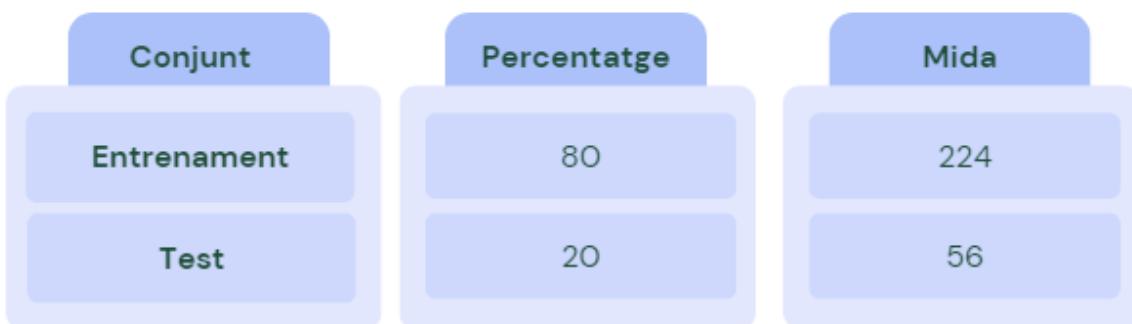


Figura 55: Particionat

## 2.6 Tractament de missings

El primer pas per tractar els valors faltants o també anomenats missings, va ser detectar-los. A la figura 77 es pot veure la quantitat de missings que tenen les variables.

Per detectar-los, com que hi havia diferents estils de missings: 'nan', un espai en blanc,, 'NaN', per exemple, vaig començar convertint-los tots en un únic estil de missing, en aquest cas, vaig utilitzar el tipus de missings de pandas. D'aquesta manera, vaig simplificar la feina a l'hora de trobar els missings. Observant la notable quantitat de valors faltants a la variable 'drug' que representa aproximadament el 25% de les dades de la variable, vaig decidir no gestionar aquests missings i eliminar les files que contenen missings en aquesta variable.

Aquesta decisió la vaig prendre després d'analitzar el significat de la variable, és a dir, determinar si el pacient ha pres placebo o un medicament específic. Vaig considerar que predir aquesta variable pot resultar en molts errors ja que no només influeixen altres estats físics tractats en la base de dades com el nivell de colesterol o coure en sang sinó que també es pot veure afectada per un factor psicològic.

En segon lloc, vaig tenir en compte també la proporció dels valors faltants. El fet de tenir tants

Atribut	Missing values	Atribut	Missing values
ID	0	Bilirubin	0
N_Days	0	Cholesterol	106
Status	0	Albumin	0
Drug	105	Copper	106
Age	0	Alk_Phosphatase	106
Sex	0	SGOT	106
Ascites	105	Tryglicerides	106
Hepatomegaly	105	Platelets	7
Spiders	105	Prothrombin	2
Edema	0	Stage	6

Figura 56: Valors de missings

missings podia afectar la fiabilitat dels mètodes de tractament de missings. I finalment, vaig reforçar la meva idea llegint la informació sobre la base de dades, on en un preprocessament que es va realitzar, van decidir eliminar les files amb valors mancats a la columna de 'drug'.

Així doncs, després d'analitzar-ho, vaig optar per eliminar les files corresponents a les dades amb valors faltants a la variable 'drug'. Aquesta decisió no només simplifica el conjunt de dades sinó que també garanteix que les observacions amb valors faltants a la variable no contribueixen amb informació incorrecta o imprecisa a l'anàlisi. Malgrat la pèrdua d'alguna informació, crec que aquesta aproximació és la més apropiada. A més a més, és rellevant destacar, que les files amb valors faltants a la columna 'drug', també tenien altres valors faltants a la columna 'drug' també presentaven valors mancats a les altres columnes. Aquest aspecte es va fer evident quan vaig eliminar les columnes, i els valors faltants de les altres variables es van reduir significativament fins a només quedar valors missings en quatre variables numèriques: la variable cholesterol, la variable tryglicerides, la variable copper i la variable platelets.

Un cop eliminats els missings de la variable de drug i de la resta de variables categòriques, vaig decidir aplicar knn amb k=5 per imputar les variables numèriques que encara contenien missings. Per fer-ho, vaig utilitzar l'algorisme KNNImputer de sklearn i vaig entrenar el model amb les dades del train i vaig utilitzar el model per imputar tant en el test com en el train.

Atribut	Missing values
Cholesterol	22
Copper	2
Tryglicerides	24
Platelets	3

Figura 57: Valors de missings sense els missings a 'Drug'

Finalment, un cop tractats els valors missings de la meva base de dades, vaig voler tornar a analitzar les distribucions per veure si aquestes es veien afectades. Tan sols vaig mirar la distribució de les variables imputades amb KNN, ja que la distribució de les altres en principi s'ha de mantenir igual.

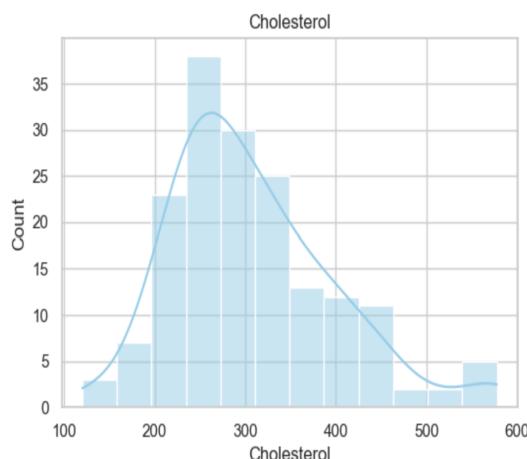


Figura 58: Distribució cholesterol sense imputació

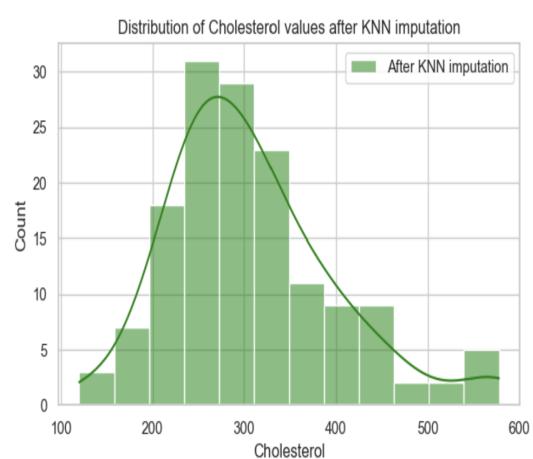


Figura 59: Distribució cholesterol amb imputació

Figura 60: Distribucions cholesterol

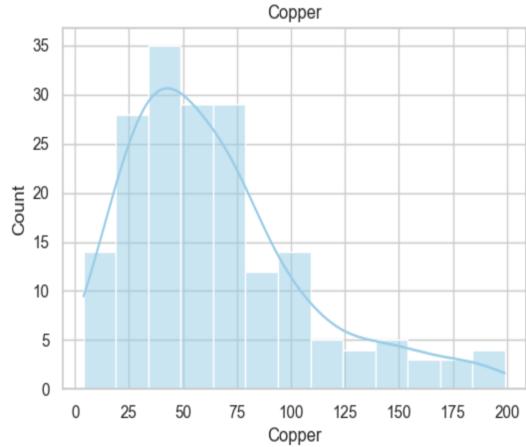


Figura 61: Distribució copper sense imputació

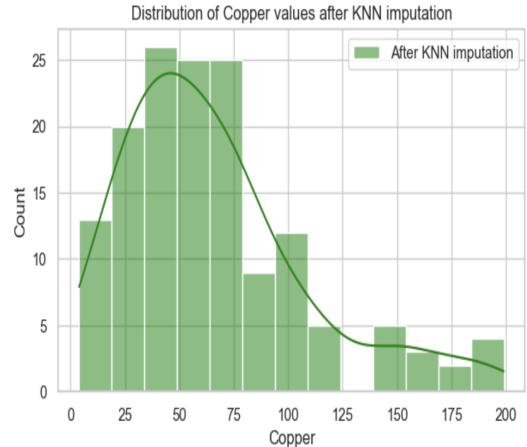


Figura 62: Distribució copper amb imputació

Figura 63: Distribucions copper

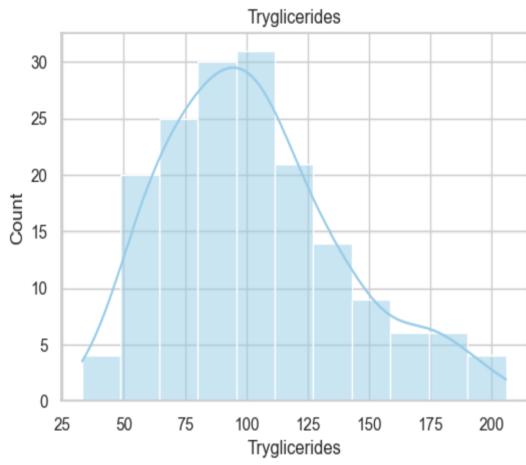


Figura 64: Distribució tryglicerides sense imputació

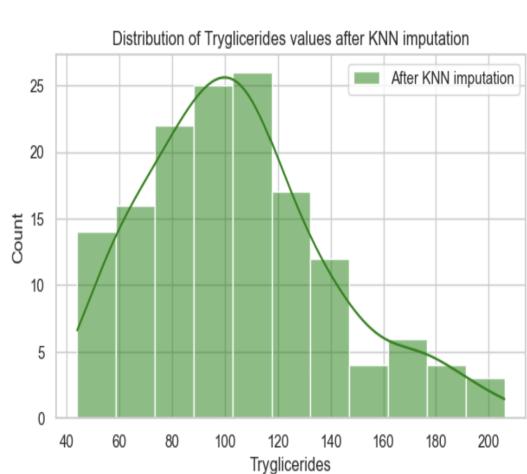


Figura 65: Distribució tryglicerides amb imputació

Figura 66: Distribucions tryglicerides

En aquest cas. podem veure que la imputació ha mantingut les distribucions en els quatre casos. Podem dir que l'única que té un canvi és a la variable Platelets. No obstant això, aquest canvi és mínim.

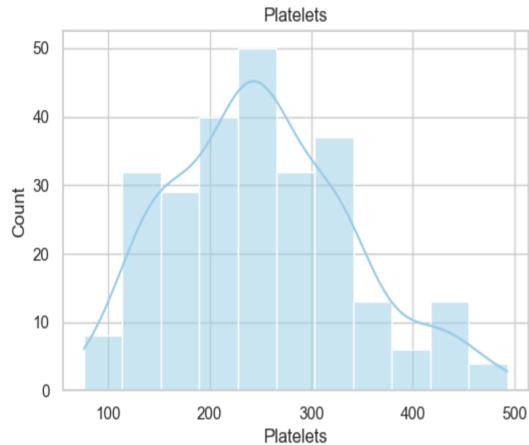


Figura 67: Distribució platelets sense imputació

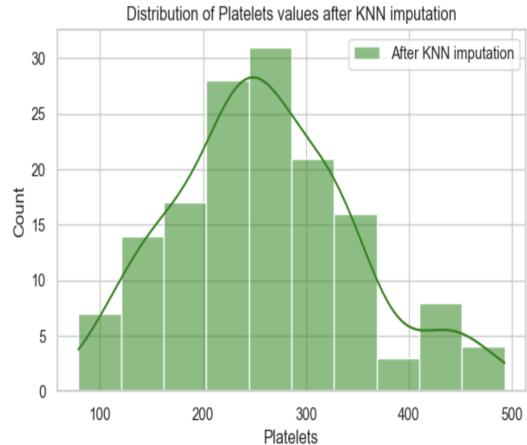


Figura 68: Distribució platelets amb imputació

Figura 69: Distribucions Platelets

## 2.7 Tractament del desbalanceig de classes

Després de realitzar l'anàlisi univariant de les variables, vaig poder identificar un desbalanceig significatiu en la classe objectiu, coneguda com a 'status'. A més a més, també vaig detectar desbalanceig en les variables 'sex', 'ascites', 'spiders', 'edema' i 'stage'.

Per abordar aquest desequilibri, vaig optar per aplicar una tècnica de sobre-mostratge (overfitting), ja que no disposava de suficients dades per utilitzar una tècnica de sub-mostratge (underfitting). Per fer-ho, vaig aplicar l'algorisme SMOTE, centrant-me especialment en la classe objectiu. Un cop aplicat el mètode de balanceig, vaig tornar a estudiar les distribucions de totes les categoriques.

### Drug:

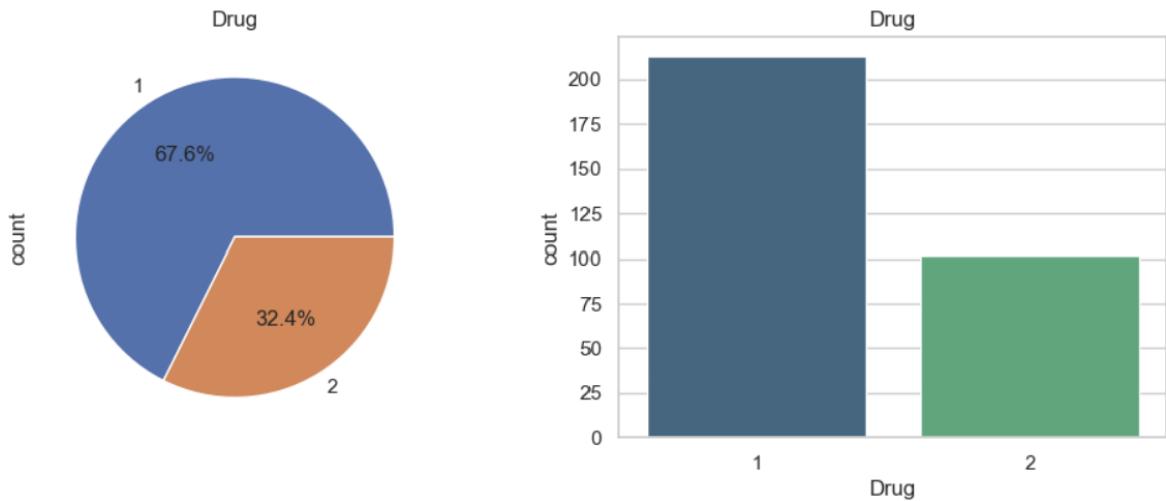


Figura 70: Variable drug amb el balanceig

En aquest cas podem observar com la variable ha obtingut un desbalanceig important ja que

anteriorment, tenia una proporció del 49.4% de la classe 1, corresponent al placebo i un 50.6% per la classe 2, corresponent al D-penicillamine.

#### **Sex:**

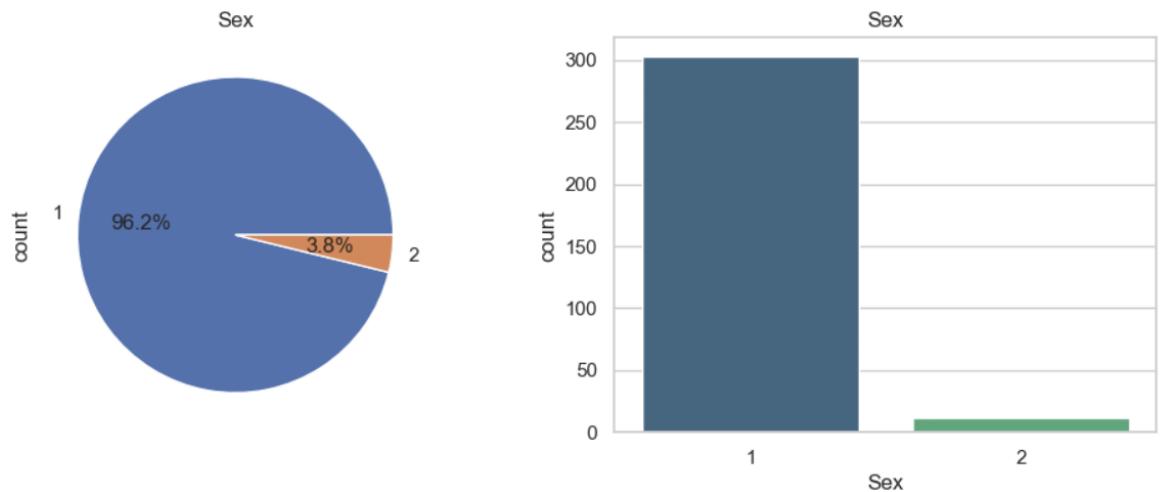


Figura 71: Variable sex amb el balanceig

En aquest cas, també veiem com la variable s'ha desbalancejat més que abans. Abans d'aplicar el balanceig, la classe 1 corresponent a F tenia un percentatge de 89.5% mentre que la classe 2, corresponent a M, comptava amb una representació del 10.5%.

#### **Ascites:**

El mateix passa amb la variable ascites, on anteriorment el 92.3% dels valors pertanyien a la classe 1, corresponent a N, mentre que un 7.7% pertanyien a la classe 2, corresponent a 2.

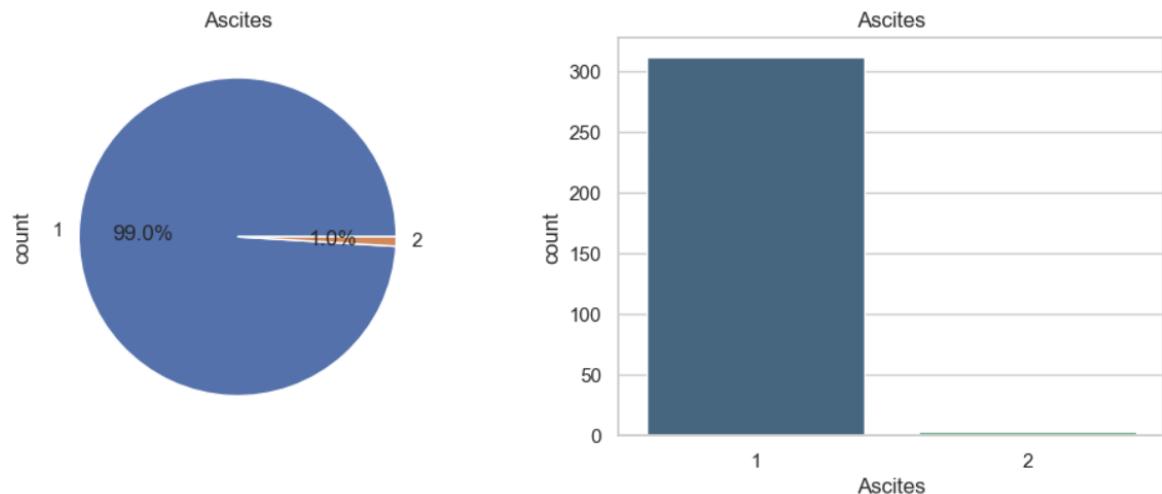


Figura 72: Variable ascites amb el balanceig

### Hepatomegaly:

En aquest cas, tornem a tenir un desbalanceig evident ja que anteriorment la classe 1 comptava amb un 51.3% de la representació, sent la classe Y. Per altra banda, la classe 2 (N), ocupava el 48.7% dels valors.

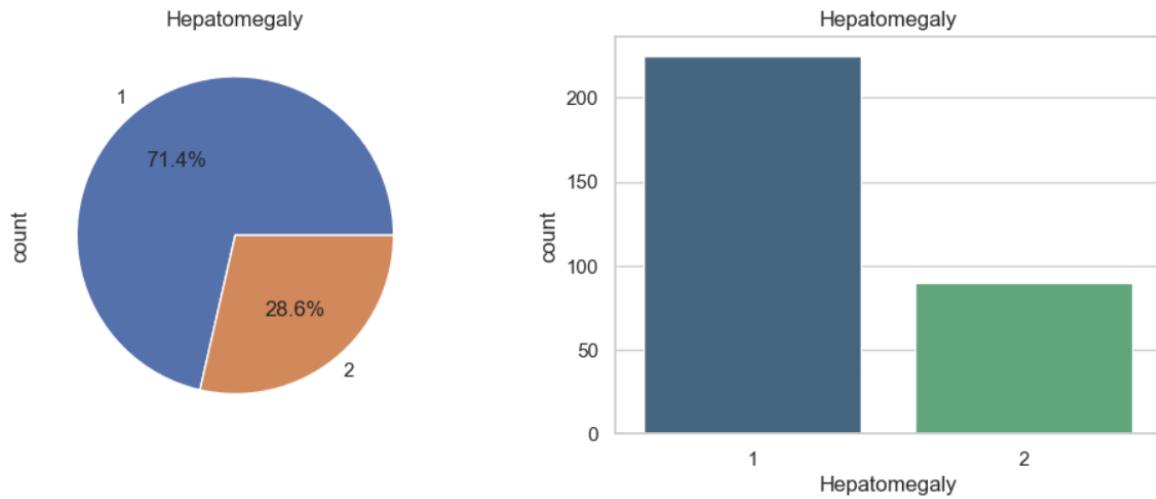


Figura 73: Variable hepatomegaly amb el balanceig

### Spiders:

En el cas de la variable spiders, podem dir com la variable s'ha balancejat bastant, ja que inicialment, la classe 1, corresponent a Y, comptava amb el 28.8% dels resultats, mentre que la classe 2, corresponent a N, ocupava el 71.2% dels valors.

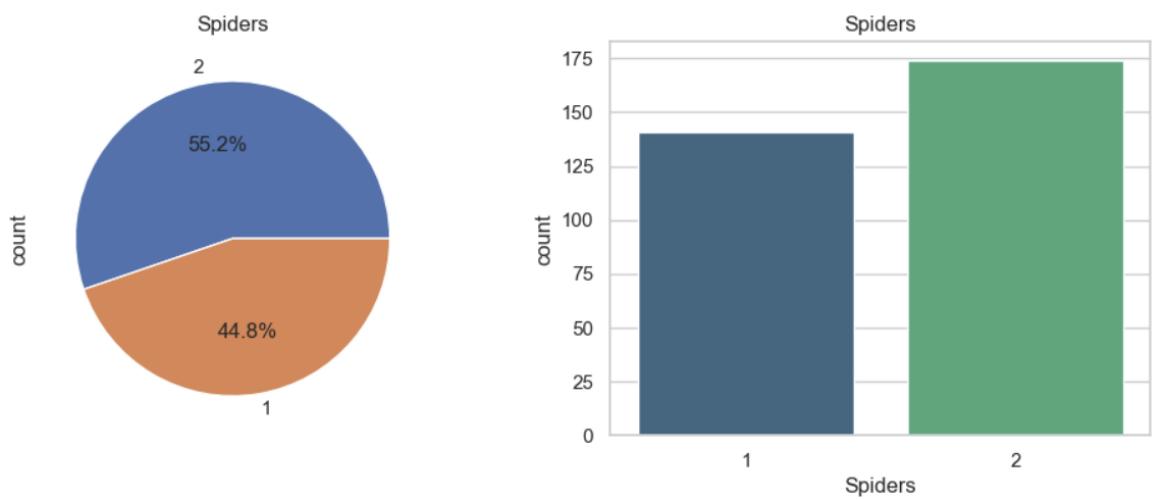


Figura 74: Variable spiders amb el balanceig

### **Edema:**

En el cas de l'Edema, encara s'ha desbalancejat més fent molt poc probables la classe 2 i 3, corresponent a S i Y. En el cas anterior, a la classe 1, corresponent a N, tenia el 84.7% dels valors mentre que S (classe 2) tenia el 10.5% dels valors i Y (classe 3) el 4.8% dels valors.

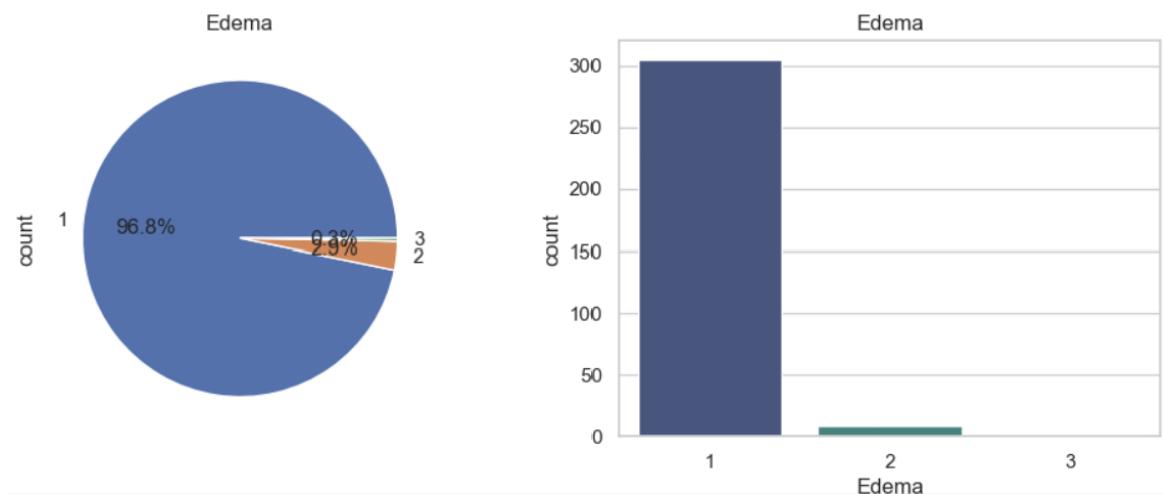


Figura 75: Variable Edema amb el balanceig

### **Stage:**

En el cas de la variable stage, la classe 1 que era la menys freqüent amb un 5.1% dels valors, ara és menys freqüent encara. La classe 2 ha augmentat la representació (anteriorment tenia un 22.3%) mentre que el valor 3 ha augmentat, ja que abans tenia un 37.6% de la representació. Finalment, podem apreciar com la classe 4 ha perdut representació, ja que anteriorment, comptava amb un 35% de la representació.

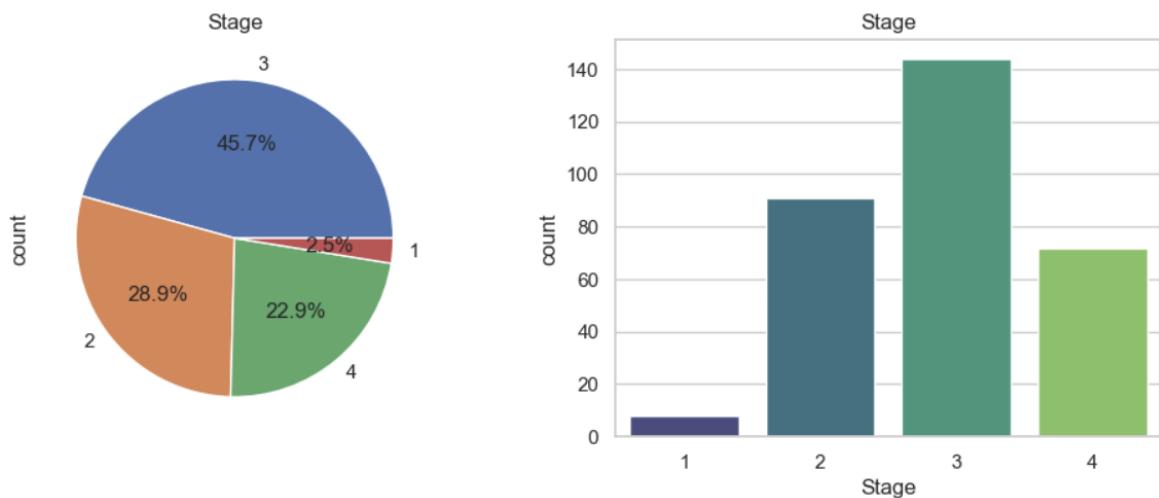


Figura 76: Variable stage amb el balanceig

### Status:

Finalment, podem veure com la variable objectiu, i la variable que m'interessava tenir balancejada, s'ha balancejat correctament. No obstant això, més endavant estudiaré altres mètodes per balancejar ja que aquest ha suposat un desbalanceig important per la majoria de variables.

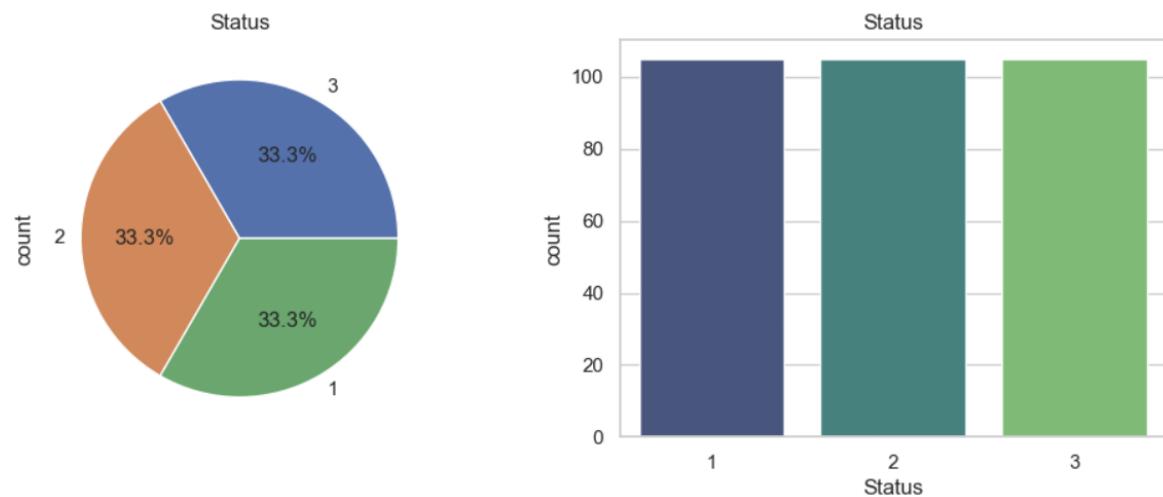


Figura 77: Variable status amb el balanceig

## 3 Preparació de variables

### 3.1 Normalització

La normalització de les dades és un pas crucial per garantir que totes les variables tinguin la mateixa escala. Aquest fet pot millorar significativament el rendiment d'alguns algoritmes, especialment els que es basen en distàncies o magnitud entre les característiques.

En el cas del KNN, la normalització és essencial, ja que l'algoritme calcula les similituds utilitzant distàncies, i si les dades no estan normalitzades, les característiques amb valors més grans podrien dominar la distància total. La normalització assegura que totes les característiques contribueixin de manera equitativa a la distància.

L'algoritme SVM, també es basa en distàncies, és per aquesta raó que aquest algoritme també recal normalització.

En canvi, els arbres de decisió no requereixen normalització, ja que es basen en comparacions relatives entre característiques i no en les magnituds absolutes.

Pel que fa a l'algoritme EBM, malgrat ser robust enfront de les diferències d'escala, la normalització encara pot ser beneficiosa, millorant l'eficiència de l'algoritme.

Després de considerar aquestes raons, vaig decidir aplicar la normalització com a pas final en el meu model. Vaig optar per aplicar el mètode Min-Max, ja que vaig considerar interessant mantenir la distribució de les dades després de la normalització. Un cop aplicada aquesta tècnica, vaig examinar les dades normalitzades per analitzar-ne la seva distribució.

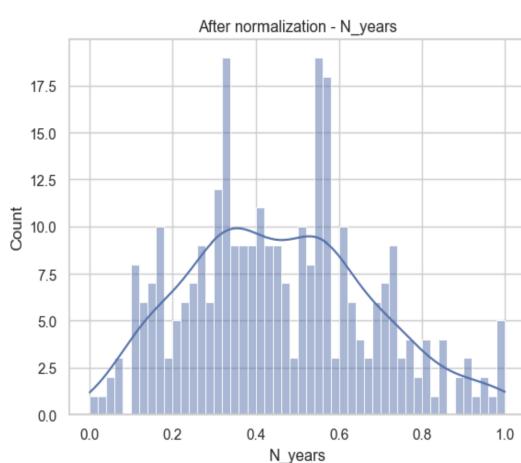


Figura 78: Distribució N\_Years normalitzada

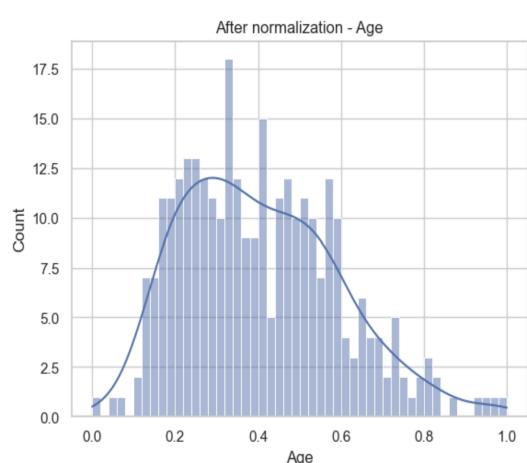


Figura 79: Distribució Age normalitzada

Figura 80: Distribucions N\_years i age normalitzades

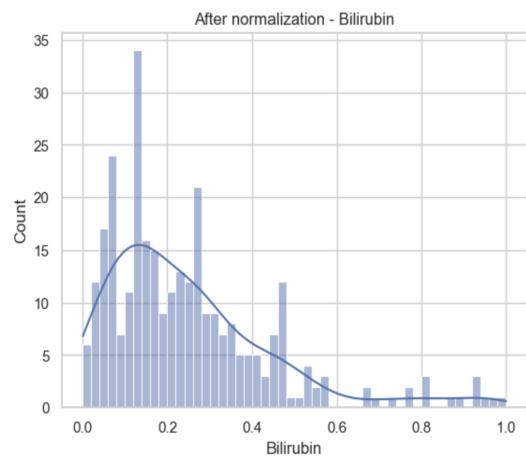


Figura 81: Distribució bilirubin normalitzada

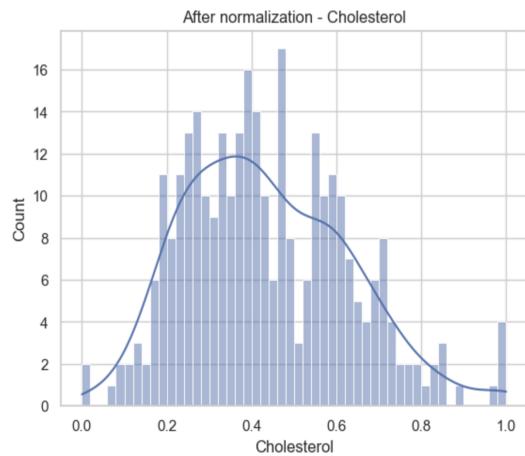


Figura 82: Distribució cholesterol normalitzada

Figura 83: Distribucions Bilirubin i Cholesterol normalitzades

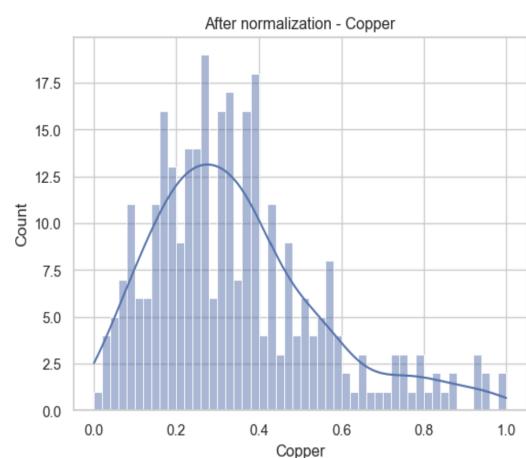


Figura 84: Distribució Copper normalitzada

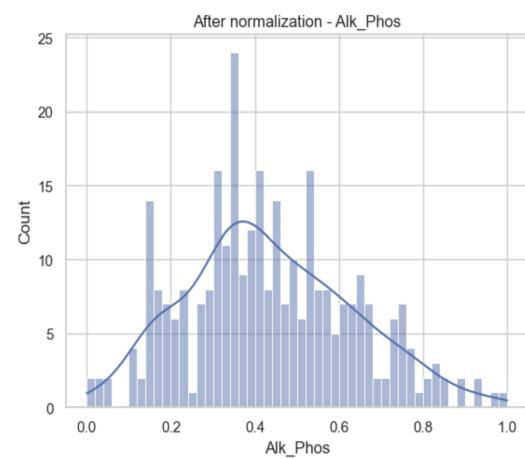


Figura 85: Distribució Alk\_Phosphatase normalitzada

Figura 86: Distribucions Copper i Alk\_Phosphatase normalitzades

Observant les noves distribucions de les gràfiques després d'aplicar la normalització, podem observar com els valors ara es troben continguts en un rang de 0 a 1. Comparant les noves distribucions amb les distribucions de les gràfiques sense normalitzar, es pot observar com la majoria de les variables mantenen la seva forma original de distribució. Això significa que, tot i haver ajustat els valors a una escala específica, la informació relativa a la variabilitat i la relació entre els diferents punts de dades es preserva en gran mesura.

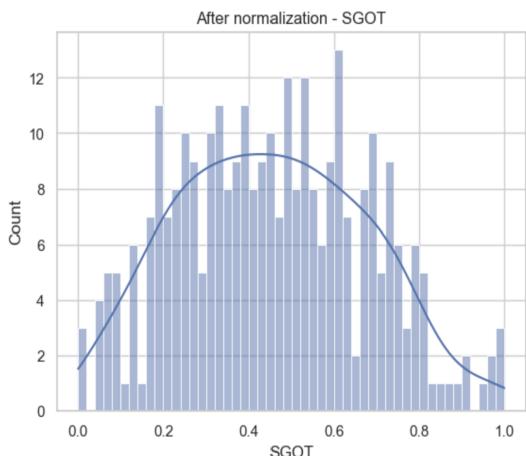


Figura 87: Distribució SGOT normalitzada

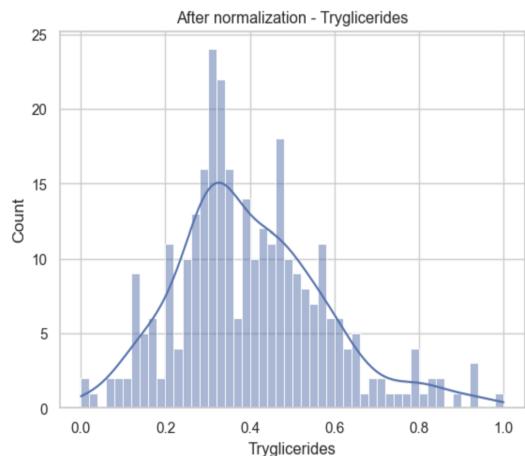


Figura 88: Distribució Tryglicerides normalitzada

Figura 89: Distribucions SGOT i Tryglicerides normalitzades

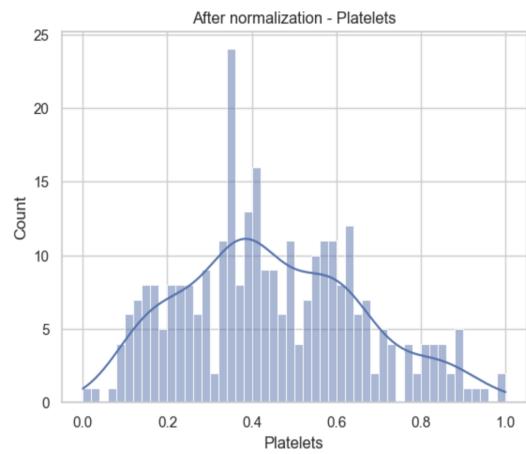


Figura 90: Distribució Platelets normalitzada

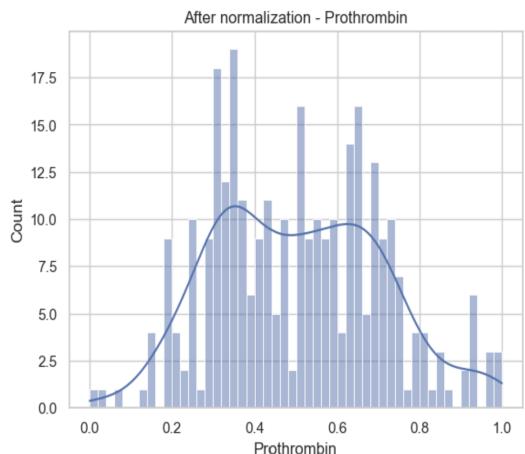


Figura 91: Distribució Prothrombin normalitzada

Figura 92: Distribucions Platelets i Prothrombin normalitzades

### 3.2 Anàlisi de correlacions

Per aprofundir en la comprensió del conjunt de dades, vaig realitzar una anàlisi de correlacions entre les variables numèriques. Aquesta anàlisi em va permetre comprendre les relacions entre aquestes variables i poder detectar redundància entre les variables, és a dir, variables que proporcionaven informació similar a altres variables. En aquest cas, podria eliminar les variables sobrants per disminuir la dimensió de la base de dades.

En la imatge 93 podem observar una matriu de correlacions per totes les variables. En aquest cas, sense aplicar la imatge es fa difícil comprendre els eixos, no obstant això, només cal saber que l'ordre de variables tant per l'eix X com l'eix Y és: ID, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phosphatase, SGOT, Tryglicerides, Platelets, Prothrombin.

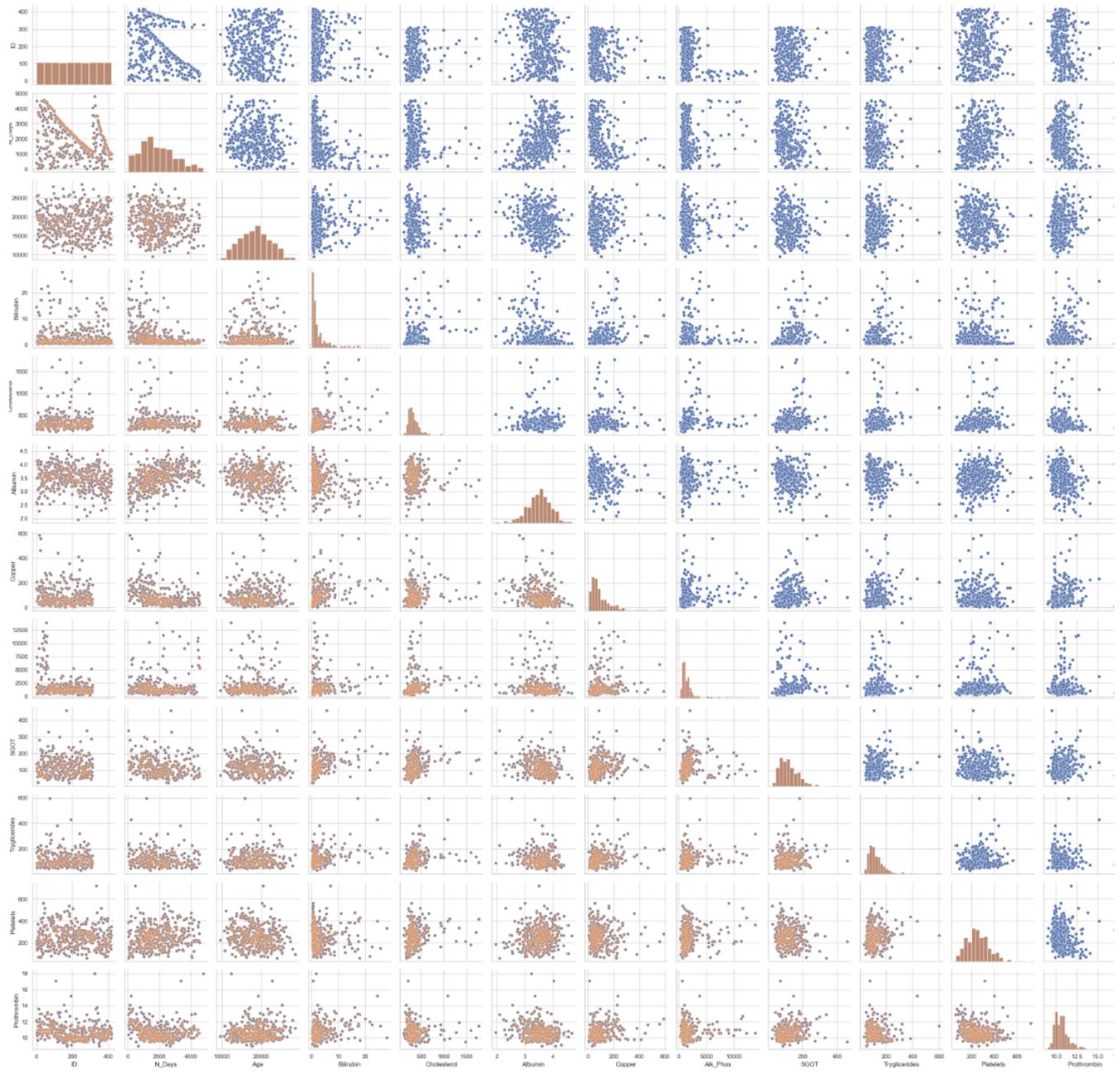


Figura 93: Anàlisi correlacions

Per facilitar l'anàlisi, també vaig utilitzar la matriu de correlacions següent:

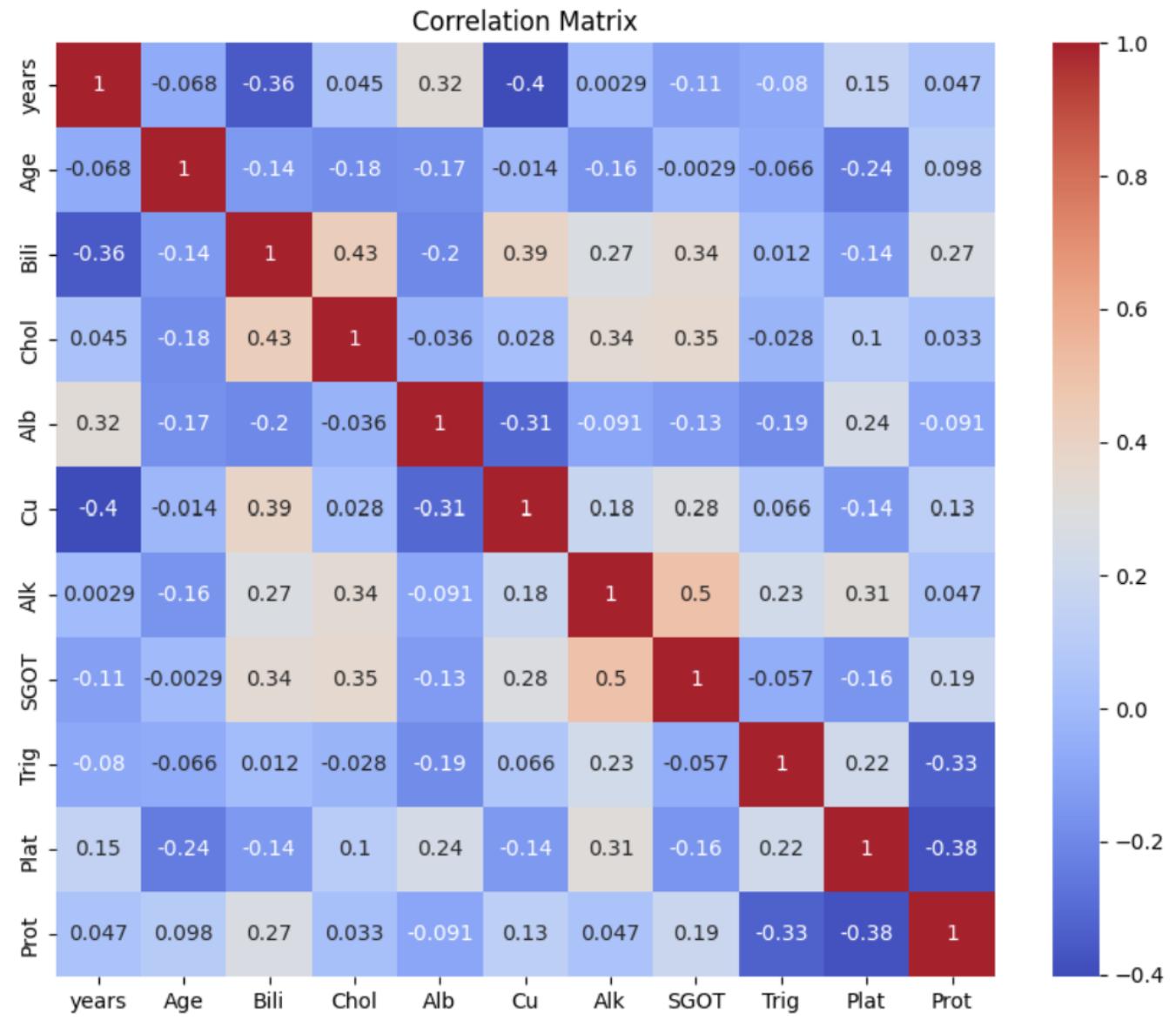


Figura 94: Anàlisi correlacions

La matriu anterior mostra com es relaciona cada variable amb les altres. Els valors pròxims a 1 o -1 indiquen una correlació forta, ja sigui positiva o negativa, mentre que valors pròxims al 0 indiquen que no hi ha correlació o que n'hi ha molt poca.

Observant la matriu, no es veuen correlacions extremadament altres, el que suggereix, tot i que hi ha correlacions moderades i fortes (a baix nivell).

En aquest cas, podem observar que les relacions més altes es troben entre SGOT i Alk\_Phosph. Aquestes dues variables presenten una correlació positiva forta, del 0.5, el que ens pot sugerir que hi ha una associació significativa entre ells. No obstant això, aquesta correlació no és suficientment gran

per poder eliminar una de les dues variables.

També trobem una correlació mitjana del 0.43 entre les variables colesterol i bilirubina, això pot indicar que alts nivells de bilirubina poden estar associats amb alts nivells de colesterol.

Analitzant les relacions negatives, trobem una relació negativa del -0.36 entre la variable edat i bilirubina, indicant que a mesura que l'edat augmenta, els nivells de bilirubina tendeixen a disminuir, o viceversa.

També trobem una relació negativa entre l'Albumina i el coure. En aquest cas, la correlació és moderada (-0.31) indicant les majors concentracions de coure poden estar associades a nivells d'albumina menors.

Un cas similar passa entre els triglicèrids i les plaquetes que en aquest cas, tenen una correlació de -0.33, indicant que alts nivells de triglicèrids estan associats amb baixos nombre de plaquetes, o viceversa.

En el cas del prothrombin i les plaquetes, aquesta correlació és més forta (-0.38) indicant que a mesura que els nivells de prothrombin augmenten, el nombre de plaquetes tendeix a disminuir.

Finalment, podem trobar variales que no tenen correlació o que aquesta és molt baixa, com en el cas del colesterol i els anys, o els anys i Alk\_phos o l'edat i SGOT, entre d'altres.

Aquestes són unes quantes conclusions que podem extreure després d'analitzar les matrius de correlacions entre les variables. No obstant això, aquestes matrius m'interessen sobretot per poder detectar les variables irrelevants. En aquest cas, la correlació més elevada es troba a 0.5, indicant que malgrat que aquestes variables tenen una correlació forta, aquesta no és suficientment forta com per dir que una de les dues variables és irrelevant.

Seguidament, també vaig voler analitzar cada variable amb la variable objectiu. En aquest cas,

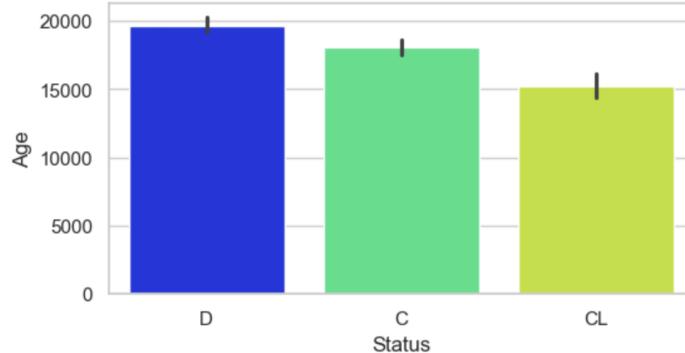


Figura 95: Anàlisi Age

podem veure que la categoria 'D' (death), és la que té l'edat més elevada. Per altra banda, la categoria 'CL' (sobreviure amb trasplant) és la més baixa del tot però no per molt en comparació amb la categoria 'C' (sobreviure). En aquest cas es pot observar com la categoria de mort, té valors de

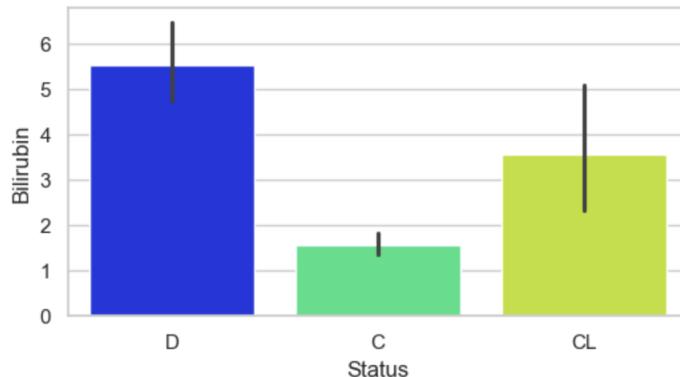


Figura 96: Anàlisi Bilirubin

bilirubin més elevats mentre que la categoria amb uns nivells de bilirubin menors és la categoria 'C' (sobreviure).

En el cas de la figura 97, podem veure com no hi ha molta diferència en els nivells de colesterol entre les categories. No obstant això, la categoria amb un nivell de colesterol més elevat, i amb una major variabilitat, és la categoria 'CL', mentre que la categoria 'C', tenen els valors més baixos de colesterol.

En el cas de la figura 98, podem apreciar com no hi ha un canvi significatiu entre les tres classes i els nivells d'albumin del pacient.

En el cas de la figura 99, podem veure com les categories amb valors de coure més elevats són les categories 'D' i 'CL', aquesta última amb una variabilitat molt elevada.

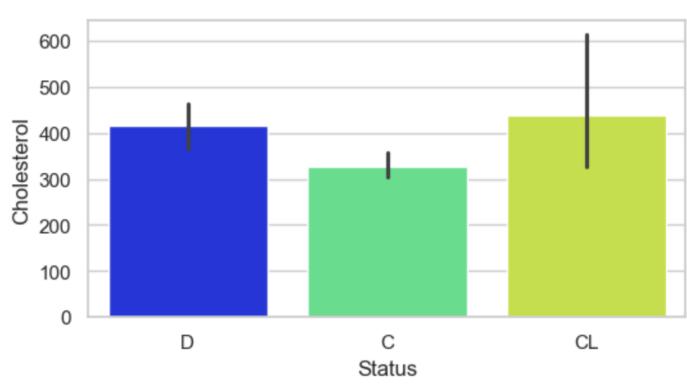


Figura 97: Anàlisi Cholesterol

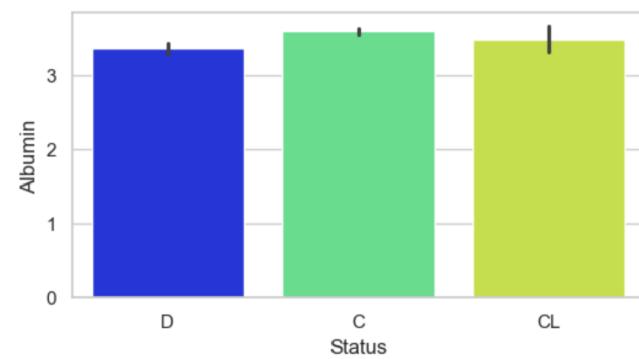


Figura 98: Anàlisi Albumin

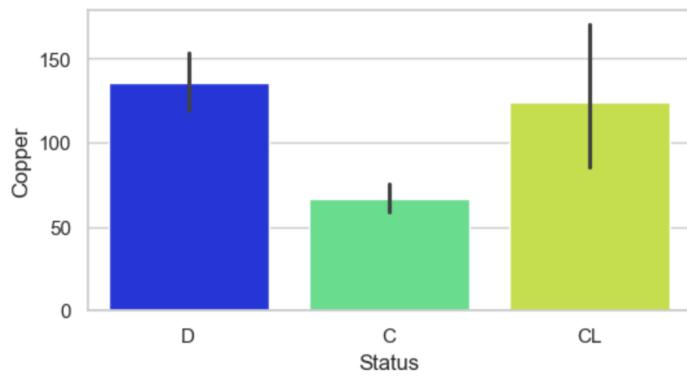


Figura 99: Anàlisi Copper

En el cas de la variable Alk\_Phosphat, els valors més elevats es troben en la categoria 'D' mentre que les altres dues categories tenen valors extremadament semblants.

En el cas de la figura 101, podem apreciar com els valors de SGOT més elevats es troben en la classe 'D' i 'CL'. No obstant això, la classe 'C', té uns valors una mica més baixos que les altres classes. El mateix passa amb la variable Tryglicerides.

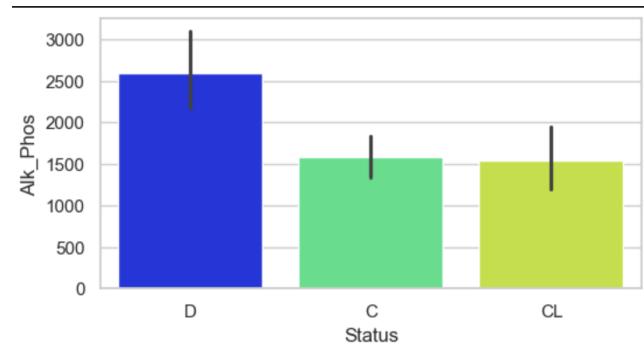


Figura 100: Anàlisi Alk\_Phos

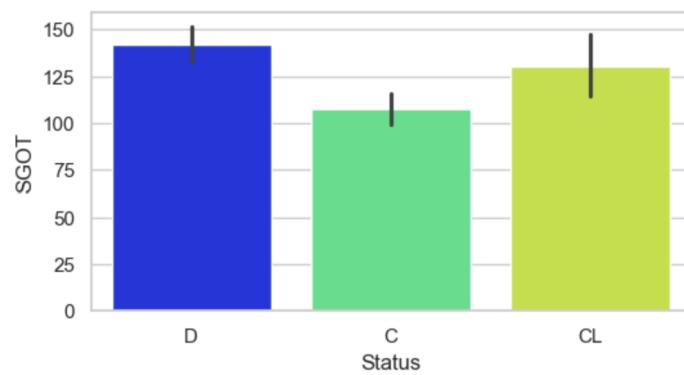


Figura 101: Anàlisi SGOT

En el cas de la variable Platelets, els valors més elevats es troben a la categoria 'CL' mentre que els valors més baixos, amb una diferència poc notable, es troben en la classe 'D'.

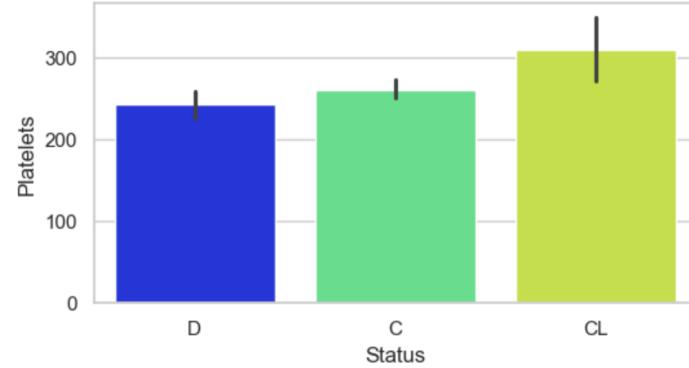


Figura 102: Anàlisi Platelets

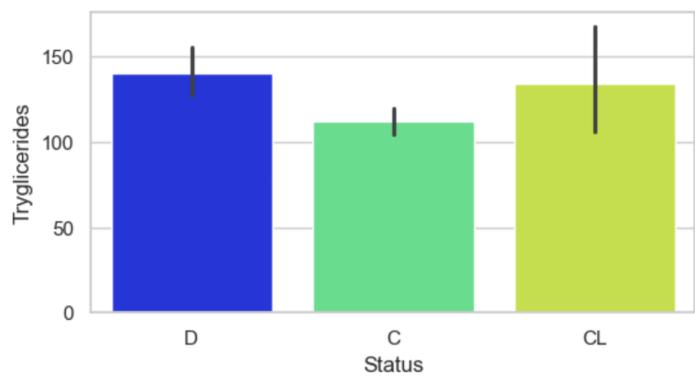


Figura 103: Anàlisi Tyglicerides

Finalment, observant la variable prothrombin, podem notar una diferència poc significativa entre les diferents classes.

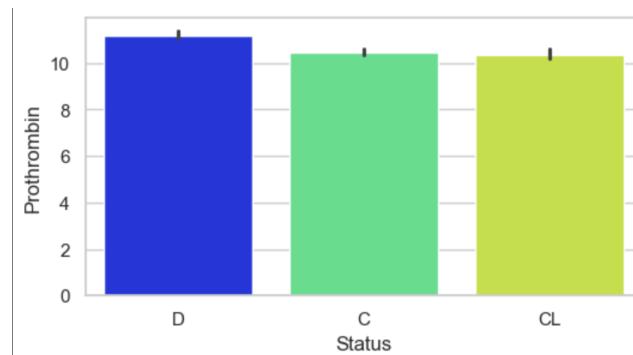


Figura 104: Anàlisi Prothrombin

### 3.3 Anàlisi variables categòriques

Per analitzar les variables categòriques, ho vaig fer analitzant al mateix instant, com es comporta la variable a estudiar amb la variable objectiu.

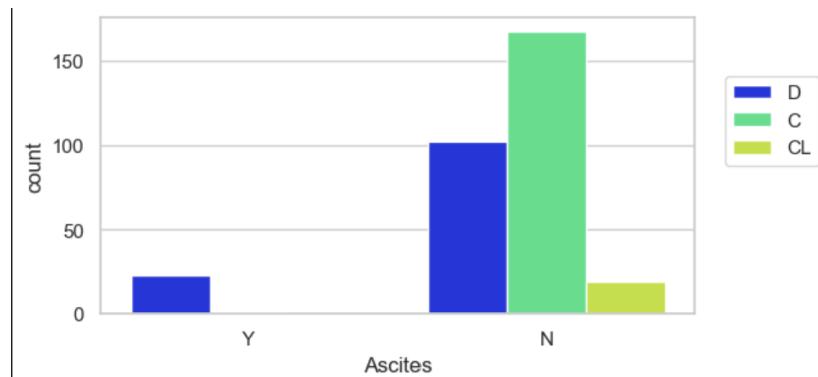


Figura 105: Anàlisi ascites

En el cas de les Ascites, podem observar com els casos que presenten ascites són molt menors als que no en presenten, de fet, trobem casos no representatius com que no hi ha pacients que hagin sobreviscut amb ascites, sigui amb trasplantament o sense.

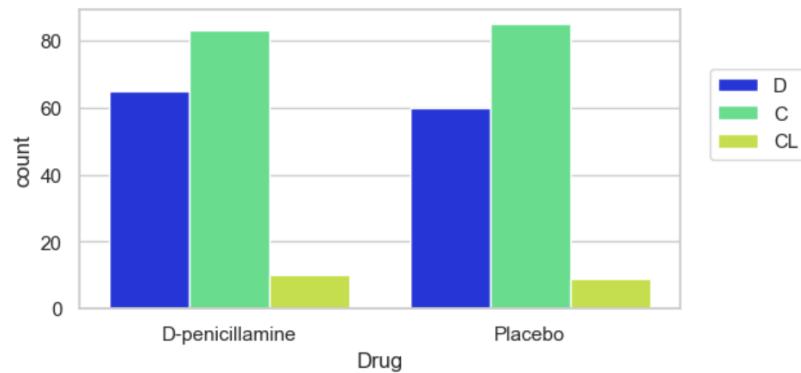


Figura 106: Anàlisi drug

Per la variable drug, veiem un comportament similar tan si s'ha pres D-penicillamine o placebo. Només es pot apreciar una diferència entre els valors de les persones que han mort, que és superior pels clients que han pres D-penicillamine.

Amb el cas de l'Edema ens trobem en un cas similar a l'ascites, on un percentatge molt elevat de persones no presenten edema, el que fa que no hi hagi representació en alguns casos com les persones que tenen edema i sobreviuen.

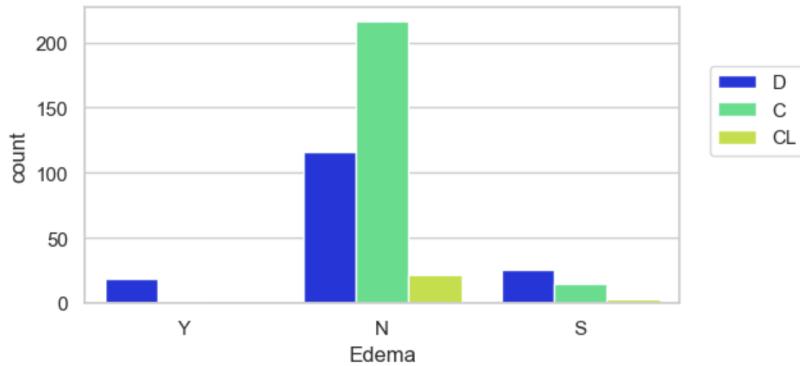


Figura 107: Anàlisi Edema

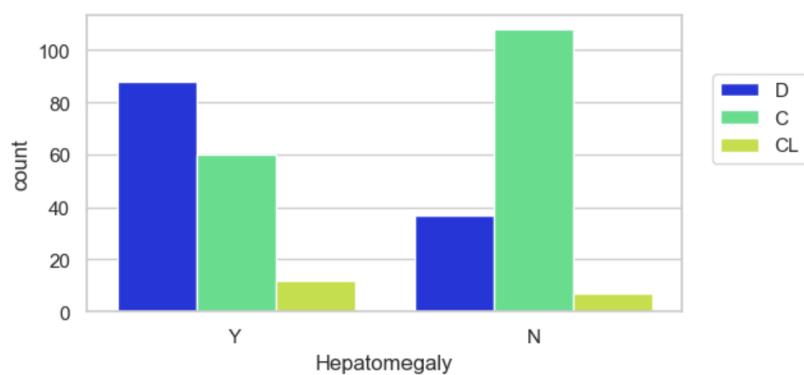


Figura 108: Anàlisi Hepatomegaly

En el cas de la variable Hepatomegaly, podem observar com hi ha més gent que en té i que mor que no pas les que en tenen i sobreviuen. Per altra banda, hi ha més gent que en té i sobreviu respecte la gent que en té i no sobreviu.

Per la variable del sexe, veiem que hi ha una poca representació del sexe femení, el que fa que

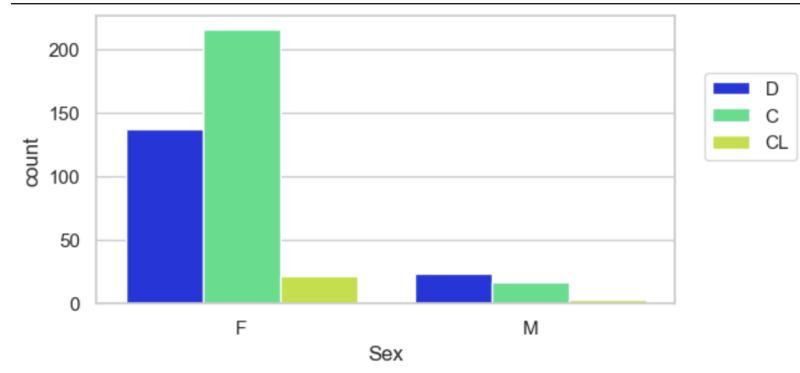


Figura 109: Anàlisi sex

no hi hagi cap representació de persona masculina que hagi sobreviscut amb trasplant.

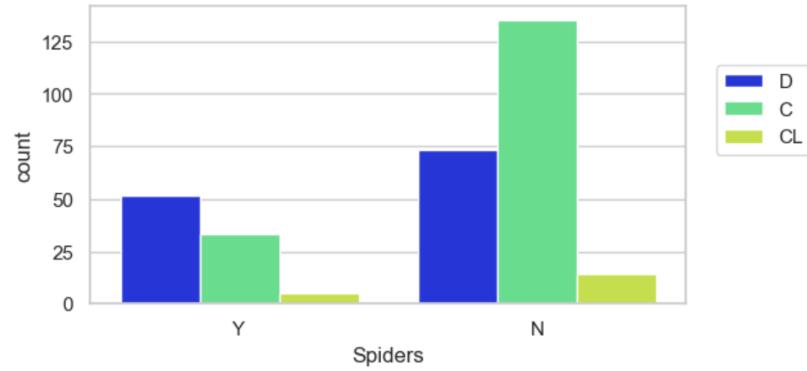


Figura 110: Anàlisi spiders

En el cas de spiders, podem observar com hi ha més persones que no tenen spiders que sobreviuen, mentre que en el cas de les personnes que tenen spiders, gran quantitat mor. No obstant això, trobem un desbalanceig important entre les dues classes. Finalment, observant la matriu de la figura 167

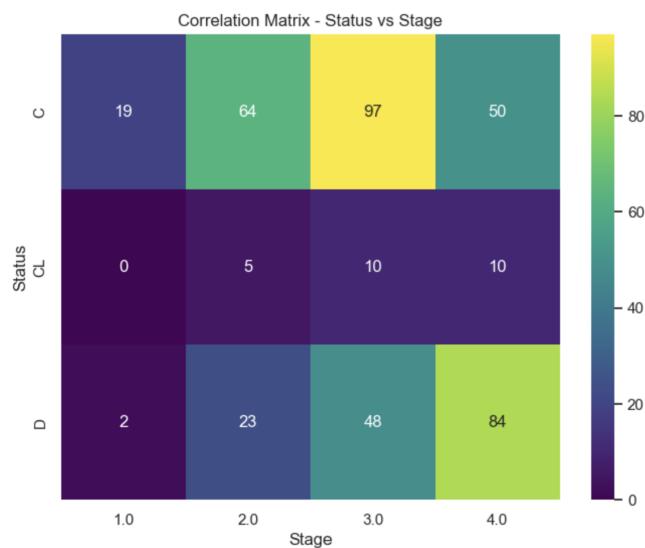


Figura 111: Anàlisi stage

podem adonar-nos que en general hi ha una baixa representació de la categoria 'CL' mentre que trobem una gran quantitat de valors per aquelles persones en estat 2 i 3 que es recuperen i de persones que estan en l'estat 3 i 4 i moren.

### 3.4 Estudi dimensionalitat amb PCA

Seguidament, he realitzat un estudi de dimensionalitat utilitzant l'anàlisi de components principals (PCA) amb l'objectiu de comprendre la necessitat de reduir les variables en el nostre conjunt de dades, buscant així una disminució de la dimensionalitat.

En una primera fase, es va explorar el nivell de varianci explicada per cada dimensió resultant. La Figura 112 ofereix una representació visual d'aquesta variancia. Observem que la primera dimensió explica aproximadament un 24% de la variancia total, seguida de prop per la segona dimensió amb un 19%. A partir d'aquí, la variancia explicada per cada dimensió disminueix progressivament, amb la tercera dimensió representant aproximadament el 12%. Cal destacar que el nostre conjunt de dades original consta de 11 dimensions, les quals corresponen a les variables numèriques de la base de dades.

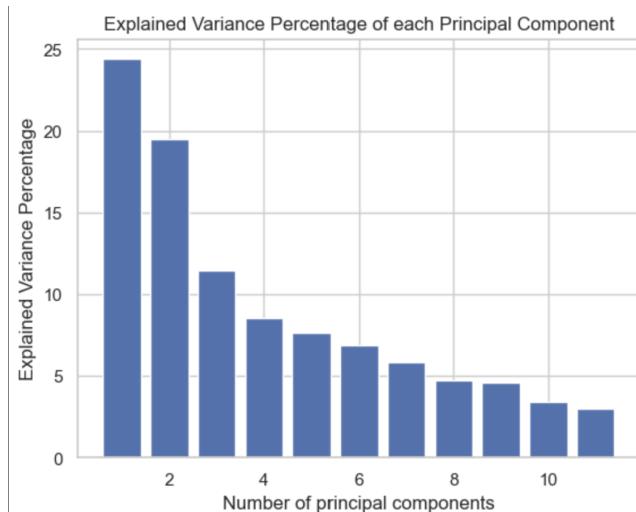


Figura 112: Anàlisi PCA

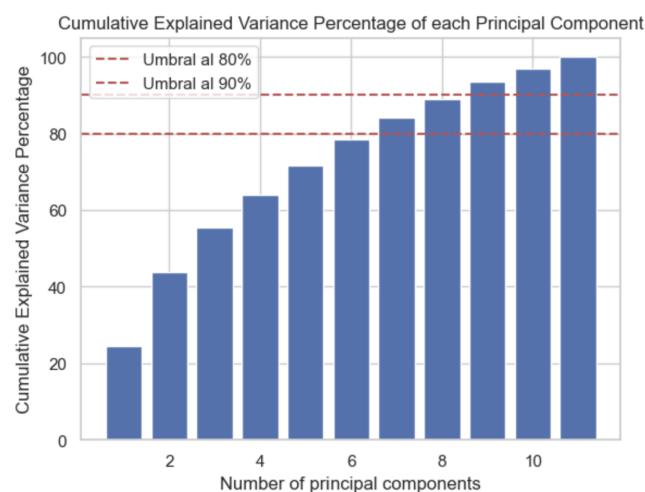


Figura 113: Anàlisi PCA

A la figura 113, es pot veure la variancia acomulada al llarg de les dimensions, mentre que a la figura 114, podem veure la variancia representada per a cada dimensió i la variancia total acomulada per les diferents dimensions.

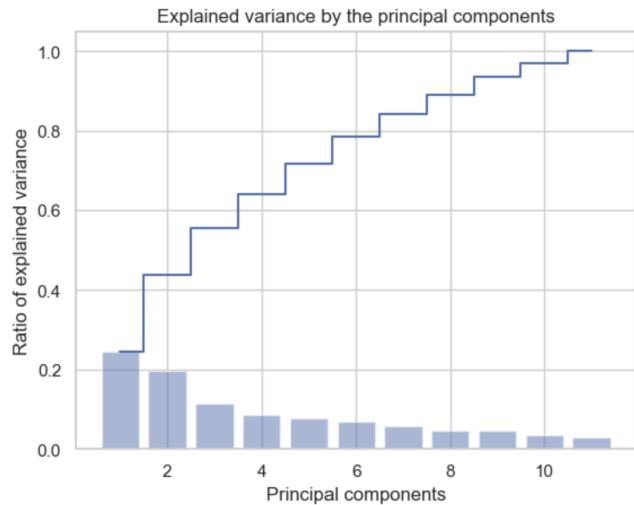


Figura 114: Anàlisi stage

En el cas de la figura 113, es poden veure dos umbrals un al 80% de la variancia total explicada i l'altre al 90% de la variancia total explicada. Marcar aquests umbrals ens ajuda a saber quantes dimensions són suficients per poder donar gairebé la mateixa informació però amb menys variables i dimensions.

És comú afirmar que una variancia del 80% ja és suficient per a molts escenaris. En aquest cas concret, aquest nivell de variancia implicaria l'ús de només 7 variables. No obstant això, com que l'objectiu principal del model és predir si una persona morirà o no, vaig considerar que aquest model té molta responsabilitat associada, i per tant, vaig optar per una variancia del 90%. Aquesta elecció té com a objectiu crear models més robusts, ja que es pretén assegurar una representació més completa del conjunt de dades original, amb l'esperança de millorar la precisió i fiabilitat dels nostres resultats predictius.

Per obtenir el 90% de la variancia, es necessita un total de 10 variables, per tant, ens proposa eliminar una variable, i conseqüentment, una dimensió.

Després de reflexionar, vaig adonar-me que eliminant tan sols una variable, no simplificava significativament la dimensionalitat del model, tot i que comportava una pèrdua d'informació. A més a més, cal tenir present el nombre limitat de dades disponibles en el conjunt, fet que fa que cada variable sigui encara més valiosa. L'impacte d'aquesta pèrdua d'informació podria ser significatiu en un conjunt de dades relativament petit.

Considerant aquestes qüestions, es va prendre la decisió de mantenir totes les variables originals, preservant així la màxima quantitat d'informació possible.

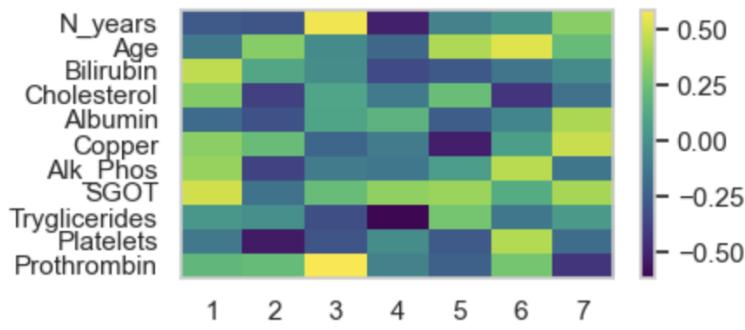


Figura 115: Anàlisi PCA

A la figura anteriorment, es pot veure un mapa de calor que mostra la correlació entre les variables i els diferents components principals obtinguts pel PCA (en aquest cas es mostren 7). Cada fila correspon a una variable, i cada columna correspon a un component principal. El color de cada cel·la indica la força i la direcció de la correlació: els tons cap al groc indiquen una correlació positiva, els tons cap al verd indicarien poca o cap correlació, i els tons cap al lila indiquen una correlació negativa.

Després d'analitzar atentament el mapa, he pogut detectar com la variable Bilirubin i Prothrombin mostren colors més intensos en diferents components principals, el que podria voler dir que aquestes dues variables tenen correlacions més fortes amb els components principals i per tant, que són significatives dins del conjunt de dades.

Per altra banda, la variable SGOT i Alk.Phos, semblen tenir una correlació positiva amb el primer component principal però una correlació negativa o neutra amb els altres components principals. Això ens pot mostrar diferents aspectes capturats pel primer component principal respecte els aspectes capturats a les altres components principals.

### 3.5 Variables sorolloses

En aquesta secció prèvia, hem estat investigant les variables per determinar si hi ha alguna redundància. Una variable clarament redundant i que s'ha eliminat és 'ID', ja que aquesta no aporta cap informació rellevant a l'individu, sinó que simplement serveix com a identificador.

En segon lloc, després d'analitzar la matriu de correlació entre les variables numèriques, he observat que cap d'elles és prou redundants, ja que els coeficients de correlació no són prou elevats. Això indica que no és necessari eliminar cap variable numèrica.

Posteriorment, mitjançant l'ús de l'anàlisi de components principals (PCA), he conclòs que no té sentit eliminar cap variable numèrica. Això és degut a que la pèrdua d'informació resultant de la reducció de dimensionalitat seria més significativa que els beneficis que podríem obtenir.

Finalment, analitzant les dades categòriques, he pogut adonar-me de que moltes de les variables afectades pel desbalanceig de la base de dades. A més a més, he pogut identificar, en algunes variables, certes relacions entre la variable resposta i les diferents categories de les variables. No obstant això, considero que degut al poc coneixement del tema, és prudent intentar conservar la major quantitat d'informació possible, ja que no tinc prou criteri per determinar si els patrons observats entre les variables són reals o simplement han sorgit a causa de les poques dades disponibles. Cal subratllar que la mida del conjunt de dades, també és una de les raons per les quals he decidit mantenir la màxima quantitat d'informació possible, ja que en aquest cas, degut a la falta de dades, tota informació que es tingui, és important.

Per tant, després de realitzar aquesta exhaustiva anàlisi, vaig acabar eliminant tan sols la variable 'ID', ja que aquesta només tenia la funció d'identificador i no suposava cap informació pels diferents individus.

## 4 Resum del preprocessament

Així doncs, un cop tractats tots aquests temes i abans d'entrar en els models creats, m'agradaria fer un petit resum dels passos que vaig realitzar per tractar les dades.

- Obtenir les dades, barrejar-les per assegurar-nos que no tinguin ordre i crear-ne copies que seran a les quals aplicarem el preprocessament
- Convertir els missings a un mateix estil
- Estudi de les variables categòriques i numèriques i les correlacions entre elles
- Recodificar les unitats de les variables 'N\_Days', 'Age' i eliminar la variable 'ID'
- Tractament outliers
- Eliminar les files amb missings a la variable 'drug'
- Label Encoding
- Fer el particionat de la base de dades
- Imputar els missings de les numèriques amb el KNN (entrenat amb les dades de train i aplicat a les dades del train i del test)
- Balancejar les dades
- Normalitzar les dades
- Estudi amb el PCA

## 5 Definició de models

Un cop preprocessades les dades, es va procedir a crear els models predictors. En total es van crear 4 models diferents explicats a continuació.

### 5.1 KNN:

El primer model creat va ser un K-Nearest Neighbors (KNN). El KNN és un algorisme conegut per la seva simplicitat conceptual ja que la seva lògica es basa en trobar els veïns més propers, la qual cosa resulta beneficiària en els problemes on la transparència del model és prioritària. A més a més, és un algorisme que pot tenir un rendiment acceptable sense la necessitat ni de grans recursos computacionals i amb un volum mitjà de dades.

Per adaptar l'algorisme a les necessitats específiques del meu problema i proporcionar una flexibilitat òptima, vaig realitzar una cerca exhaustiva entre diferents hiperparàmetres tal com es mostra a la figura següent.

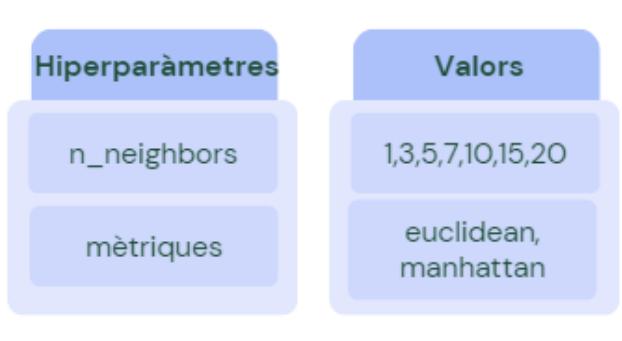


Figura 116: Hiperparàmetres KNN

En aquest context, vaig tenir en compte el nombre de veïns utilitzat per l'algorisme i la mètrica de distància, seleccionant entre les opcions euclidiana o de Manhattan.

Per determinar els millors paràmetres, vaig realitzar una recerca exhaustiva mitjançant una quadrícula i l'ús de 'GridSearchCV'. Mitjançant aquest procés i utilitzant les dades d'entrenament juntament amb diferents particions, vaig identificar els paràmetres òptims. En el nostre cas, la distància Manhattan i 1 veí. Un cop aconseguits els millors paràmetres, vaig procedir amb una validació creuada de 10-Fold per entrenar el model.

Per a cada fold, vaig enregistrar les mètriques d'accuracy, precisió, recall i F1-score. A continuació, he creat una visualització gràfica que mostra l'evolució d'aquestes mètriques durant les diferents folds. Aquesta representació proporciona una visió detallada del rendiment del model en diferents subconjunts de dades. Podem observar com les quatre mètriques del rendiment tenen una tendència similar en tots els folds, això suggereix que el model és consistent en els diferents subconjunts de dades.

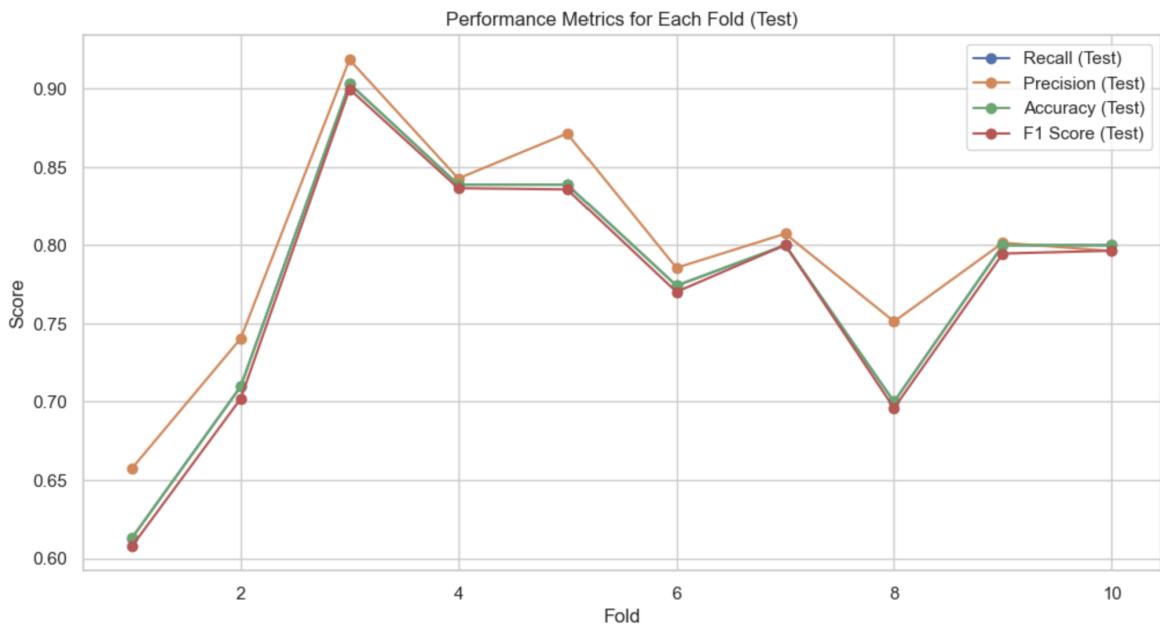


Figura 117: Resultats test de cada fold KNN

Podem observar com hi ha una variabilitat notable en el rendiment del model, això es pot deure a les diferents dades de cada subconjunt. Concretament, podem veure una baixada en el fold 7 que pot indicar un problema en les dades o en l'entrenament per causa d'aquestes.

En general, el rendiment està entre el 0.65 i 0.75 en totes les folds, el que ens indica que el model treballa bé. No obstant, la baixada a la fold 7 pot indicar que el model pateix overfitting.

Per poder analitzar millor les dades, vaig crear una altra gràfica amb l'accuracy obtinguda en l'entrenament i en el test. D'aquesta manera, podia estudiar més a fons l'overfitting. En aquesta gràfica

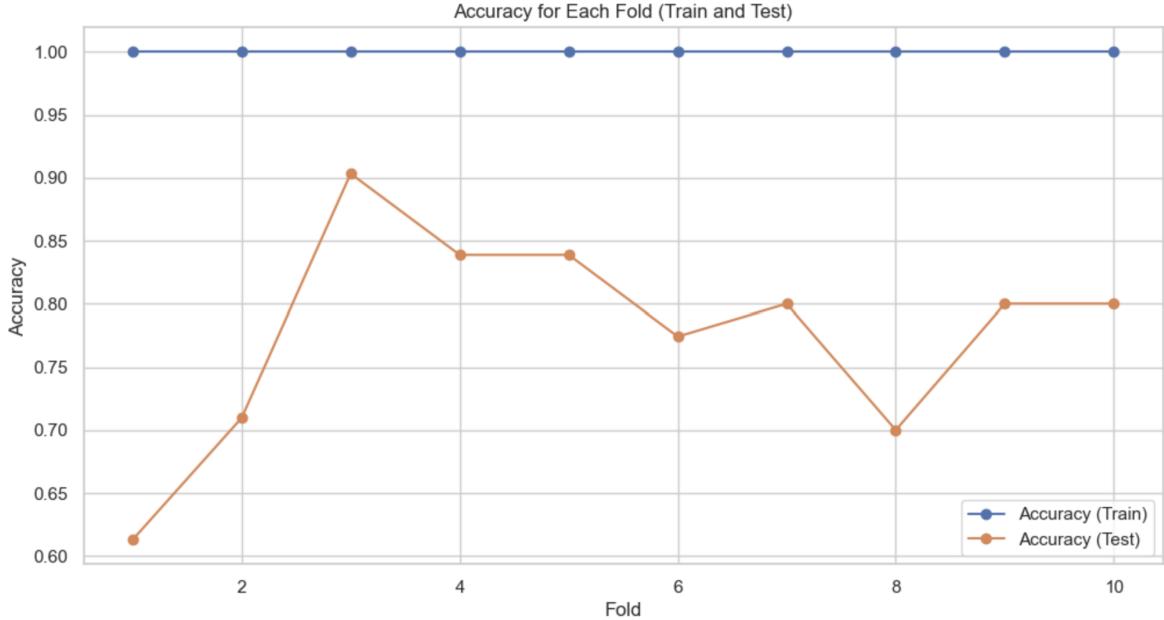


Figura 118: Resultats test i train de cada fold KNN

podem veure com l'exactitud en l'entrenament és molta alta, propera al 100% en tots els folds. El fet que hi hagi una gran diferència entre l'exactitud de les dades d'entrenament i de test pot suggerir que hi hagi sobreajustament (overfitting). Aquesta idea es reforça al fold 7 on hi ha una baixada del rendiment únicament al subconjunt d'entrenament.

No obstant això, l'exactitud del test mai és inferior a 0.65, el que indica que malgrat el possible overfitting, el model té bona capacitat per resoldre el problema amb èxit.

Finalment, també vaig crear un histograma amb la mitjana de l'accuracy del model d'entrenament i del model de test (de tots els folds). En aquest cas podem veure com l'exactitud mitjana en el

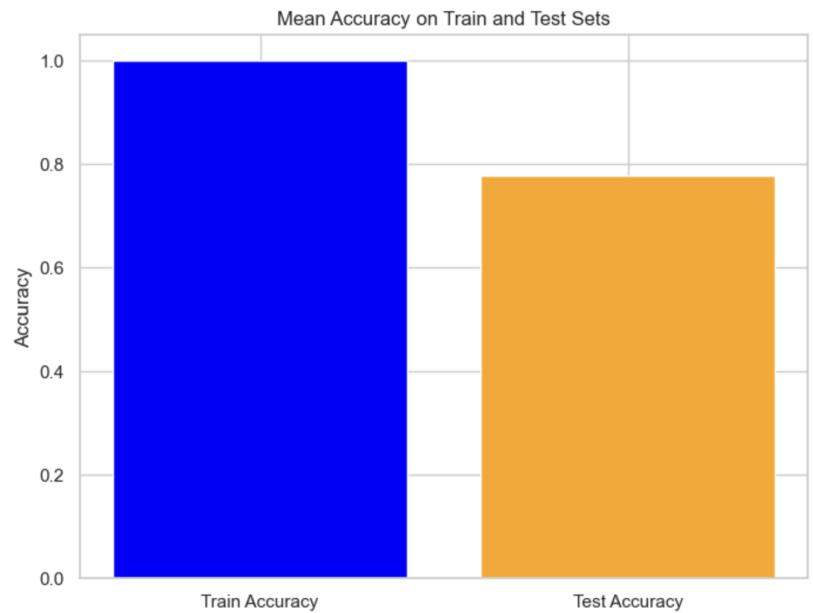


Figura 119: Resultats test i train

conjunt d'entrenament és molt alta, possiblement perfecta, ja que la barra arriba a 1.0. Això sol ser indicatiu d'un ajustament molt bo del model a les dades amb les quals s'ha entrenat. No obstant això, l'exactitud en el conjunt de dades del test és visiblement més baixa que l'exactitud d'entrenament. Això pot ser causat per un sobreajustament a les dades d'entrenament. A més a més, el rendiment en el test ens indica que hi ha un gran espai de millora de l'algorisme.

## 5.2 Arbre de decisió:

El segon model creat va ser un arbre de decisió. Els arbres de decisió són models fàcilment interpretables que prenen decisions basades en condicions simples i lògiques. En aquest cas, al tractar-se d'un context clínic, necessitem models transparents i interpretables.

A més a més, aquests models són relativament senzills d'executar i tan sols necessiten una mida moderada de dades. En aquests cas doncs, aplicar un arbre de decisió té lògica ja que és fàcil d'interpretar, té baixa complexitat i pot tractar amb un conjunt moderat de dades.

En aquest cas, a la figura següent podem analitzar els hiperparàmetres plantejats.

Hiperparàmetres	Valors
criterion	gini, entropy
min_samples_split	2,5,10
min_samples_leaf	1,2,4
max_depth	rang dels atributs de les dades
class_weight	1:3.5

Figura 120: Hiperparàmetres arbre de decisió

En aquest cas trobem el criteri de divisió (criterio). Aquest hiperparàmetre defineix l'estratègia utilitzada per mesurar la quantitat d'una divisió. En aquest cas, vaig escollir entre 'gini' per l'índex de Gini o 'entropy' per l'entropia. L'índex de Gini és més ràpid de calcular, mentre que l'entropia pot ser més sensible a les petites particions amb classes dominants.

Seguidament trobem el nombre mínim de mostres per divisió que determina el nombre mínim de mostres necessàries per dividir la base de dades.

El nombre mínim de mostres per fulla ens indica el nombre mínim de mostres que ha de tenir una fulla. Si s'augmenta aquest valor, es pot evitar que hi hagi fulles amb molt poques mostres, i per tant, es pot reduir l'overfitting.

També trobem la profunditat de l'arbre que determina la longitud màxima de camí des de l'arrel fins a les fulles. Com més profunditat, més complex és el model.

Finalment trobem l'hiperparàmetre class weight. Aquest hiperparàmetre ens ajuda a abordar situacions en que les dades del conjunt no estan balancejades. Fa referència als pesos que s'assignen a les diferents classes.

Així doncs, per determinar els millors paràmetres, vaig utilitzar, igual que en el cas del KNN, el 'GridSearchCV'. Un cop escollits els hiperparàmetres, en aquest cas un class weight de 3.5 criterion gini, max depth de 18, min leaf de 1 i min split de 2, vaig entrenar el model utilitzant un 10-fold.

A continuació podem observar els resultats dels diferents entrenaments per les diferents folds. Observant la figura anterior podem observar que les quatre mètriques mostren una tendència similar

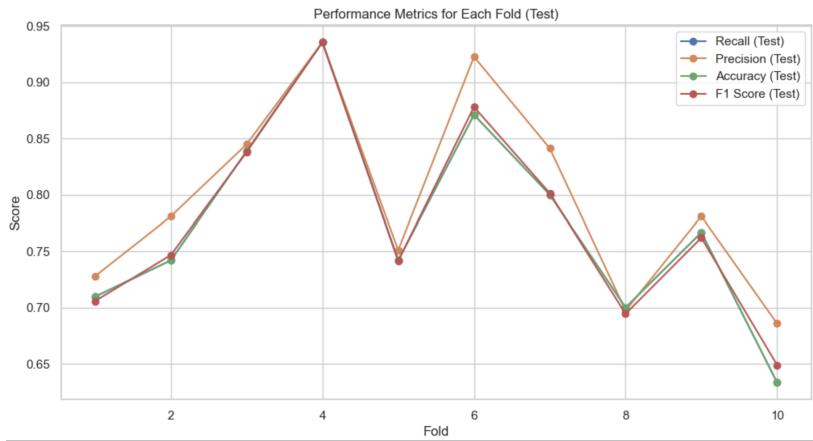


Figura 121: Resultats test de cada fold del decision tree

al llarg dels folds, el que ens indica que la variabilitat en el rendiment és consistent a través de les diferents mètriques.

En els folds 4 i 8 i 10 podem observar com hi ha una variabilitat notable en el rendiment que tendeix cap a la baixa, concretament, els valors més baixos se situen en aquests plecs. Per altra banda, els punts més elevats es troben en els folds 3 i 6, just abans de la dràstica baixada, el que pot indicar una variació en la capacitat del model per ajustar-se als diferents folds.

Finalment, podem observar com l'accuracy es manté dins d'un rang relativament elevat (entre 0.7 i 0.9), indicant que en general, el model és bastant precís.

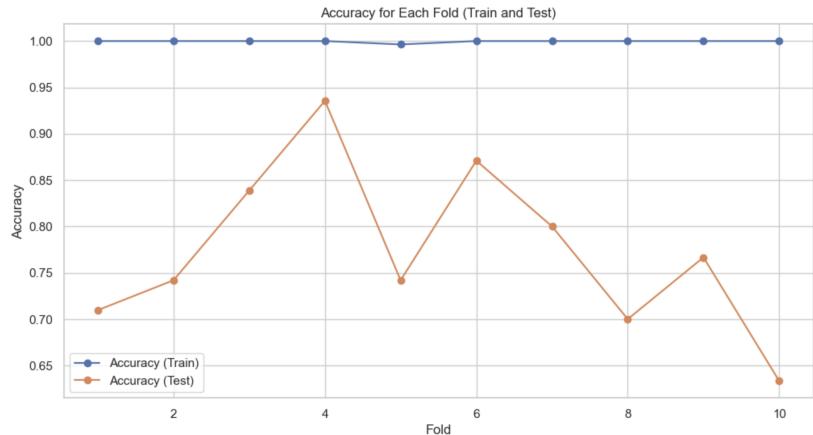


Figura 122: Resultats test i train de cada fold decision tree

En la segona gràfica podem observar com l'accuracy de l'entrenament és manté molt alta i consistent, estant a prop del 100% en tots els folds. Això ens indica que el model s'ajusta molt bé a les dades d'entrenament. No obstant això, podem observar com l'accuracy del test és relativament més baixa i conté les baixades mencionades anteriorment. Això ens pot estar indicant un sobreajustament a les dades d'entrenament, sobretot en folds 4 i 10 on l'exactitud del test és particularment baixa.

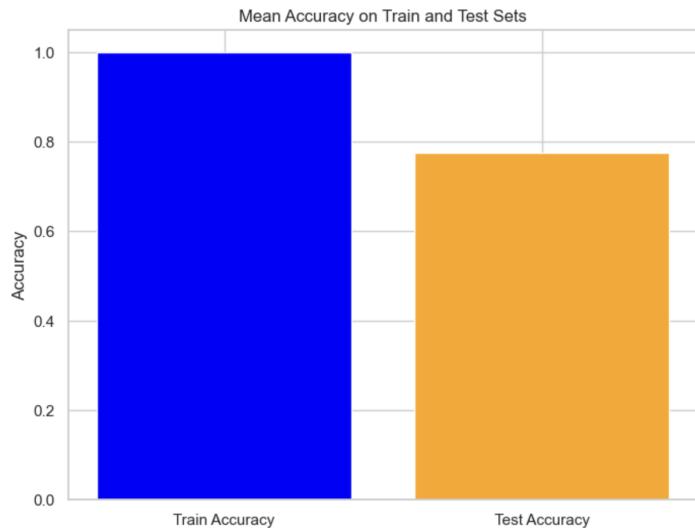


Figura 123: Resultats test i train

Finalment, en la gràfica anterior podem observar com l'accuracy d'entrenament mitjana és molt alta, tal com hem pogut observar a la gràfica anterior. Per altra banda, l'accuracy del test és més baixa que la d'entrenament, el que ens pot indicar que hi ha una generalització en el model.

### 5.3 SVM:

Seguidament, vaig crear un model de Support Vector Machines (SVM). Aquests models són adequats per tractar problemes complexos. A més a més, poden oferir una bona generalització a partir d'un conjunt d'entrenament limitat. No obstant això, són menys intuitius i difícils d'interpretar, ja que es basen en la identificació de vectors de suport en un espai de característiques.

En aquest cas, els paràmetres escollits són els següents: La C és el paràmetre de regularització

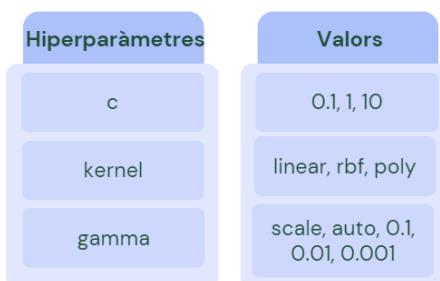


Figura 124: Paràmetres SVM

que controla la penalització per errors de classificació, és a dir, un valor elevat de C penalitzarà més els errors que no pas una C menor.

Per altra banda, el Kernel és la funció matemàtica que transforma les dades d'entrada a un espai de característiques de similitud superior. En aquest cas vaig proposar el kernel lineal, el rbf (radial basis function) i el poly (polinomial).

Finalment, la gamma controla la influència de cada mostra d'entrenament, fent que un valor petit de gamma indiqui una influència gran.

En el primer model doncs, els millors paràmetres trobats van ser 10 per la C, 0.1 per la Gamma i el kernel poly. Un cop trobats, es va realitzar el mateix entrenament que en el cas dels models anteriors. A continuació podem observar els resultats obtinguts:

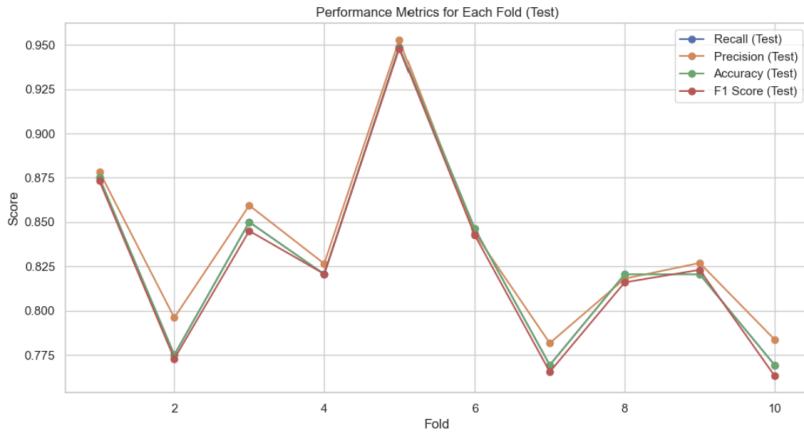


Figura 125: Rendiment SVM test

En aquest gràfica podem observar com les quatre mètriques tenen una tendència similar, per tant, tenen una consistència a través dels folds. No obstant això, podem notar una variabilitat notable especialment en els folds 2, 4, 7 i 10 on podem notar una baixada de les mètriques.

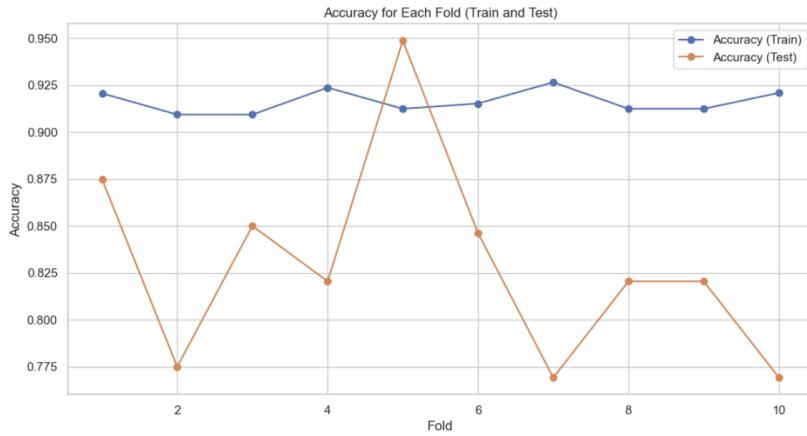


Figura 126: Rendiment SVM

Per altra banda, en aquesta segona gràfica podem observar com l'exactitud d'entrenament és consistentment alta, en els valors propers al 90% en tots els folds. No obstant això, l'exactitud del test és més variable i generalment més baixa, el que ens pot indicar un possible sobreajustament. Finalment, podem observar a l'histograma, com hi ha una baixada en el rendiment del test respecte el rendiment

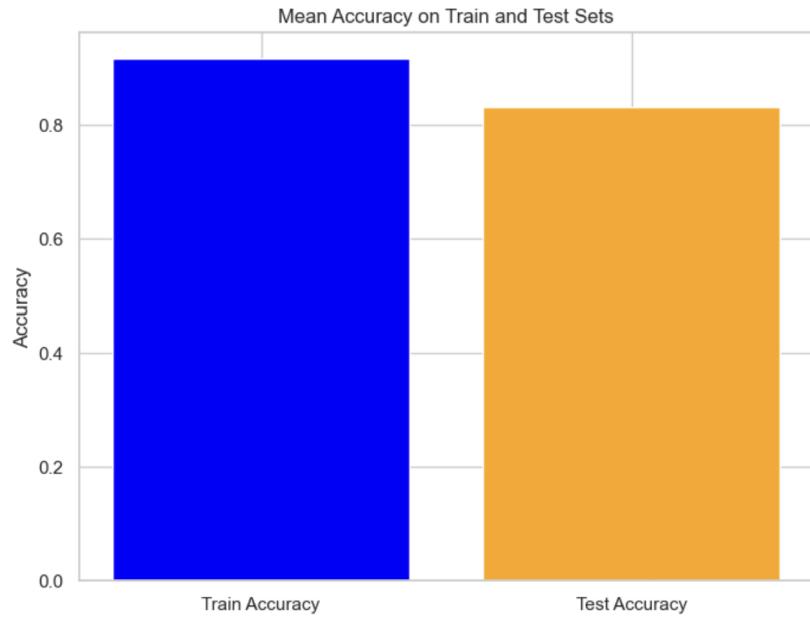


Figura 127: Resultats test i train

de l'entrenament. No obstant això, aquesta baixada no és excessivament alta, el que ens indica que el model pot resoldre correctament un percentatge important dels casos presentats.

#### 5.4 EBM

Finalment, també es va realitzar un model Explainable Boosting Machine (EBM). Aquests models són altament interpretables. A més a més, es pot adaptar a problemes de complexitat alga com a problemes més senzills i tan sols necessiten un volum de dades moderat.

A continuació, es pot observar una taula amb els hiperparàmetres provats. L'hiperparàmetre max

Hiperparàmetres	Valors
max_rounds	50,100,200
learning rate	0.01,0.1,0.2
max_bins	50,100
interactions	5,10

Figura 128: Hiperaràmentres EBM

rounds indica el nombre màxim de rondes o iteracions que l'algorisme farà durant l'entrenament. Com més rondes s'entreni, millor s'adaptarà als detalls del conjunt de dades però hi ha més risc de sobreajustament. Així doncs, no vaig voler provar valors extremadament elevats.

Per altra banda, el learning rate indica la taxa d'aprenentatge, és a dir, determina la magnitud amb què els paràmetres del model s'actualitzen a cada ronda d'entrenament. No vaig escollir nombres molt elevats ja que un valor elevat pot comportar problemes de sobreajustament.

L'hiperparàmetre Max Bins, indica el nombre màxim de caixes utilitzades per discretitzar les variables contínues. Un nombre més elevat de caixes pot proporcionar una representació més detallada de les relacions.

Finalment, el nombre d'interaccions, controla les interaccions entre variables que el model està autoritzat a aprendre.

Després de realitzar la cerca, vaig determinar que en aquest cas els millors hiperparàmetres són: un màxim de 50 rondes, un màxim de 5 iteracions, un màxim de 50 bins i un learning rate de 0.2. Seguidament, vaig entrenar el model com en els casos anteriors i vaig obtenir els següents resultats. A la primera gràfica podem observar com hi ha una inconsistència en el rendiment del model en el

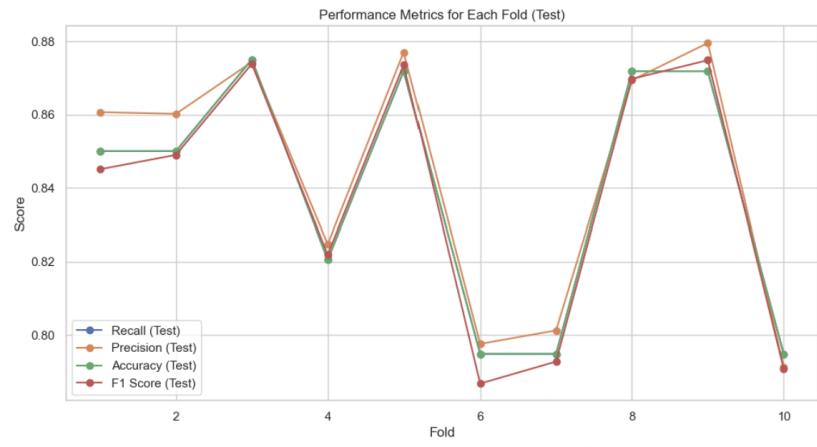


Figura 129: Resultats test i train

cas de les folds 3 i 7 on el rendiment disminueix notablement. Podem observar també com l'F1 score, que presenta el balanceig entre la precisió i el recall, varia considerablement, el que pot suggerir que el model no es manté constant entre els falsos positius i els falsos negatius.

D'aquesta gràfica en podem concloure que el model té problemes amb la distribució de les dades en folds concrets o un problema de balanceig de classes que afecta al rendiment del model. En aquest cas, podem veure com l'accuracy de l'entrenament és consistent en valors al voltant del 95%. Per altra banda, l'accuracy en el cas del test és significativament menor, sobretot en els folds 5 i 7. El que ens pot indicar que el model no generalitza bé les dades no vistes o que les dades en les folds són notablement diferents o més difícils de predir.

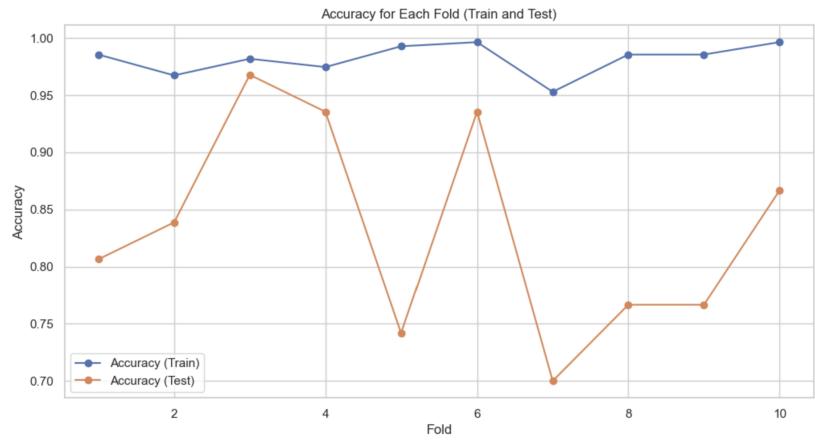


Figura 130: Resultats test i train

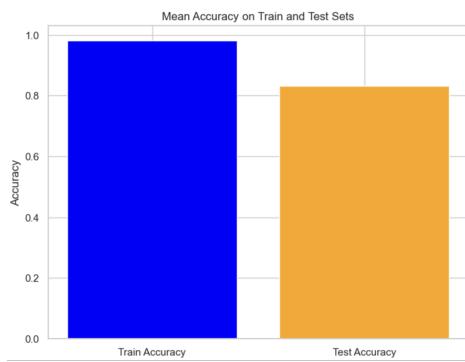


Figura 131: Resultats test i train

Finalment, a l'histograma podem observar com l'accuracy del train és superior a l'accuracy del test, tot i que aquesta última té un valor relativament alta. Això pot voler significar un nivell d'overfitting.

## 6 Selecció de models

Un cop obtinguts els models inicials, vaig decidir provar diferents maneres de preprocessar les dades per trobar el millor model.

Per tractar els outliers, vaig provar diferents rangs quartils. L'inicial era 25% i 75%, també vaig provar per 20% i 80% i per 90% i 10%. També vaig provar de passar els outliers a NA i tractar-los com a missings en comptes d'eliminar-los i de simplement deixar-los a la base de dades.

Pel tractament de missings, vaig proposar no eliminar els missings de les variables drug i aplicar el KNN per imputar els missings de les variables numèriques i la moda per les categòriques.

També vaig provar diferents mètodes per tractar el balanceig. En total vaig provar smoete, una tècnica que crea mostres sintètiques per la classe minoritària a partir dels seus veïns més propers, smotetomek que es una combinació de smote i Tomek's link i on després d'aplicar l'smote per generar les mostres sintètiques, es busquen les parelles de mostres (una de la classe majoritària i una de la classe minoritària) més properes entre si segons la mesura de Tomk's link i es remouen. I finalment, el mètode smotenc que és una extensió de l'smote però tenint en compte les variables categòriques.

I finalment, vaig provar la normalització i també l'estandardització, en aquest últim cas, es transformen les observacions perquè tinguin una mitjana de 0 i una desviació estàndard de 1.

Així doncs, per cada model, vaig provar diferents variacions en el preprocessament per finalment trobar el millor model.

A continuació podem observar les proves fetes i els models finals decidits. Per cada model, s'indiquen els paràmetres escollits i les diferents mètriques de rendiment obtingudes.

### 6.1 KNN final

En aquest cas, podem observar com el millor model, i per tant, el model escollit té un rang quartil de tractament d'outliers del 20/80%. A més a més, treballa millor si no s'eliminen els missings de la columna drug i s'imputen, les variables numèriques amb el KNN i les categòriques amb la moda. Per balancejar amb els mètodes SMOTE i estandarditzar les dades. En aquest cas, s'utilitza tan sols el veí més proper i la distància euclidiana.

			precision	recall	f1 score
		K	accuracy		
	Normalize	distance	Manhattan	0.778	0.774
	SMOTE	MinMax	Manhattan	0.758	0.752
	SMOTE	MinMax	Manhattan	0.758	0.755
	SMOTE	MinMax	Manhattan	0.758	0.755
	SMOTE	MinMax	Manhattan	0.751	0.744
	SMOTE	MinMax	Manhattan	0.753	0.748
	SMOTE	MinMax	Manhattan	0.753	0.748
	SMOTETomek	MinMax	Manhattan	0.786	0.812
	SMOTETomek	MinMax	Manhattan	0.824	0.821
	SMOTENC	MinMax	Manhattan	0.836	0.831
	SMOTENC	Standard	Euclidean	0.838	0.862
	SMOTE	Standard	Euclidean	0.856	0.864
	KNN/mode	No			0.85
Outliers	75/25	Yes	KNN		
Drop NA drugs	80/20	Yes	KNN		
Outliers	90/10	Yes	KNN		
Drop NA drugs	NA	Yes	KNN		
Outliers	-	Yes	KNN		
Drop NA drugs	80/20	Yes	KNN		
Outliers	80/20	Yes	KNN		
Drop NA drugs	80/20	Yes	KNN		
Outliers	80/20	No	KNN/mode		
Drop NA drugs	80/20	No	SMOTE		

Figura 132: Proves KNN

A continuació, podem analitzar concretament els resultats obtinguts pels diferents folds.

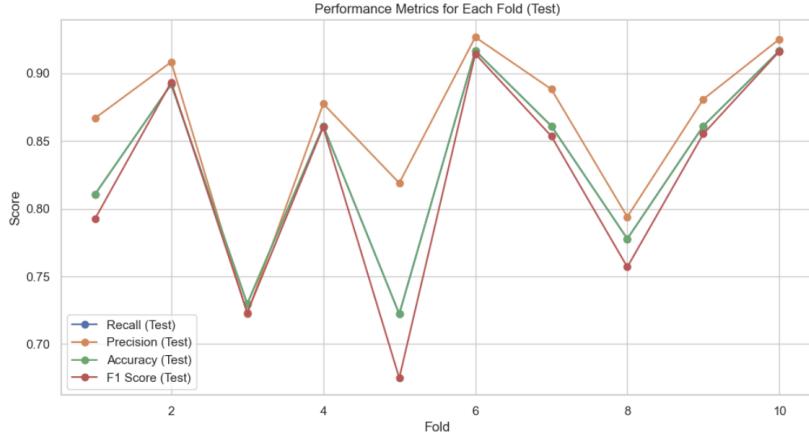


Figura 133: Model KNN final

En aquest cas, podem notar inconsistència entre els diferents folds. No obstant això, les mètriques de rendiment es mantenen en un rang consideradament elevat, entre 0.65 i 0.95 aproximadament. A més a més, podem observar com l'F1-score que és la mitjana entre la precisió i el recall és elevat el que indica un bon equilibri entre les dues mètriques i que per tant, el model no està esbiaixat ni cap a la presció ni cap al recall. Per altra banda, a la segona figura podem observar com l'accuracy

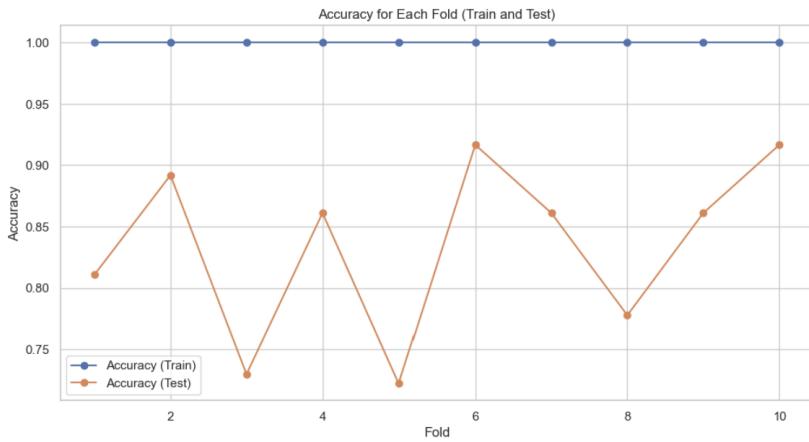


Figura 134: Model KNN final

per l'entrenament és molt alta i consistent a 1 en tots els folds, però que en el cas del test, aquesta accuracy dismíneix.

Observant la tercera gràfica podem tornar a observar l'escenari comentat anteriorment on l'accuracy de l'entrenament és major que la del test. No obstant això, aquesta diferència no és excessiva, el que podria indicar que no hi ha un sobreajustament significatiu. A més a més, el model manté una bona exactitud en dades no vistes.

## 6.2 Arbre de decisió final

En el cas de l'arbre de decisió, vaig poder veure com el millor model era aquell que elimina els outliers que no estiguin entre el 20 i 80% del rang quartil, no s'eliminen els missings de la variable drug sinó



Figura 135: Model KNN final

que s'imputen els millings numèrics amb knn i posteriorment s'aplica la moda en les categòriques. Per balancejar s'utilitza el mètode smote i s'estandarditzen les dades.

En aquest cas, els hiperparàmetres escollits són els següents: class weight = 3.5, criterion = gini, max depth = 8, min leaf = 1 i min split = 2.

Figura 136: Proves arbre de decisió

A continuació, analitzaré els resultats obtinguts. En aquest cas podem observar com el recall es

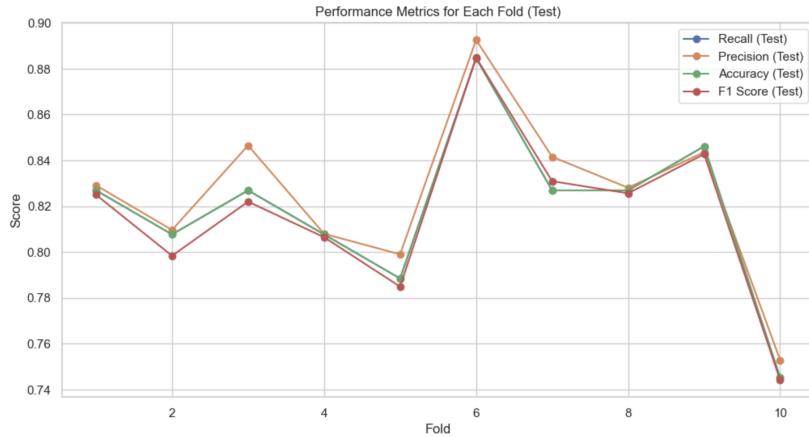


Figura 137: Model arbre de decisió final

manté relativament constant amb una tendència general lleugerament ascendent cap al final. Respecte a la precisió, aquesta mostra més variabilitat encara que no presenta canvis dràstics. L'accuracy també varia en el conjunt però sempre es manté entre 0.78 i 0.90. En general, podem veure una consistència entre les mètriques de rendiment ja que totes seguixen una mateixa tendència. Respecte a

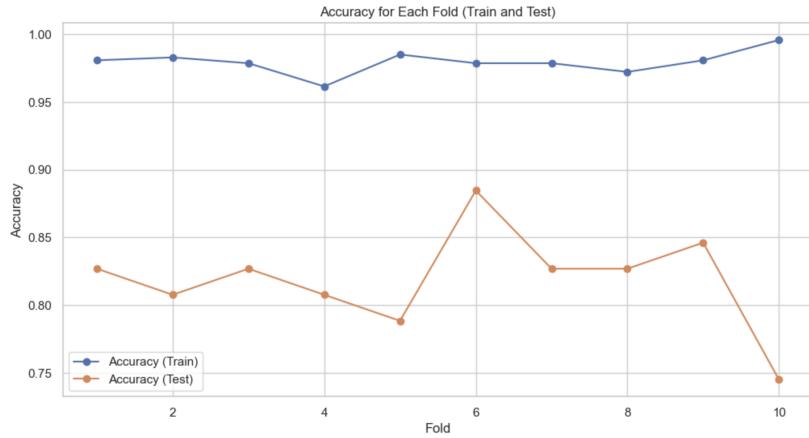


Figura 138: Model arbre de decisió final

la segona gràfica podem veure com l'exactitud de l'entrenament és alt i consistent e que indica que hi pot haver un sobreajustament, ja que en el test aquesta disminueix i mostra més variabilitat.

La tercera gràfica mostra aquesta diferència entre les exactituds (o accuracies) entre el conjunt de dades i de test, malgrat poder indicar que hi ha sobreajustament a les dades d'entrenament, el rendiment a les dades del test és alt.

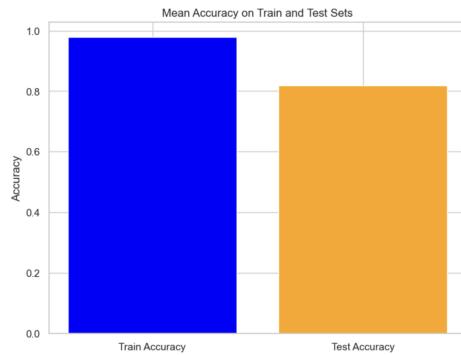


Figura 139: Model arbre de decisió final

### 6.3 SVM final

En el cas del model SVM, el model final amb el qual vaig decidir quedarme, és un model on en el preprocessat de dades no s'eliminen ni es tracten els outliers, s'eliminen els missings de la variable drug, a més a més s'aplica l'algoritme KNN per imputar i smote pel balanceig i es normalitzen les dades amb MinMax.

Respecte als hiperparàmetres, la c pren el valor de 10, la gamma és 0.1 i el kernel és rbf.

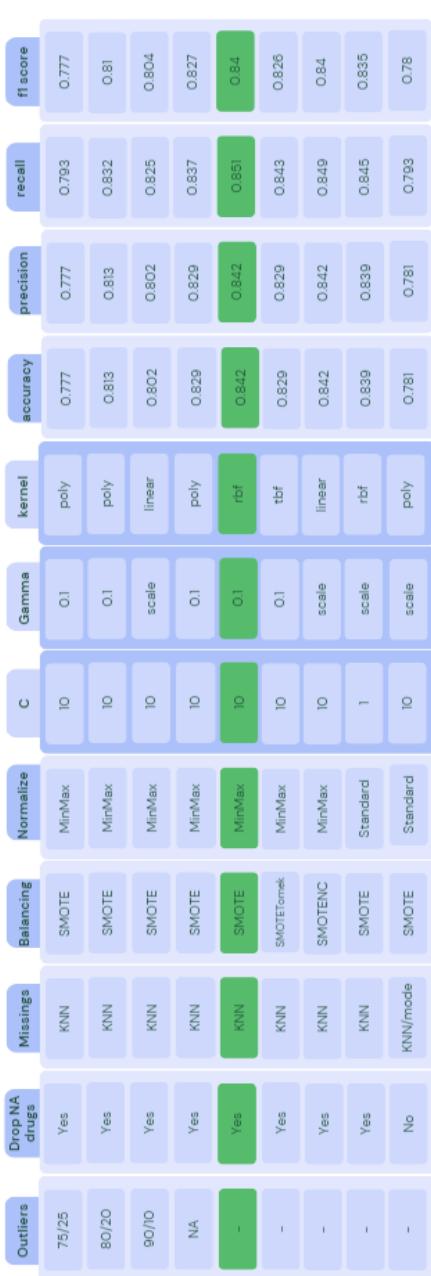


Figura 140: Proves SVM

A continuació analitzaré els resultats obtinguts.

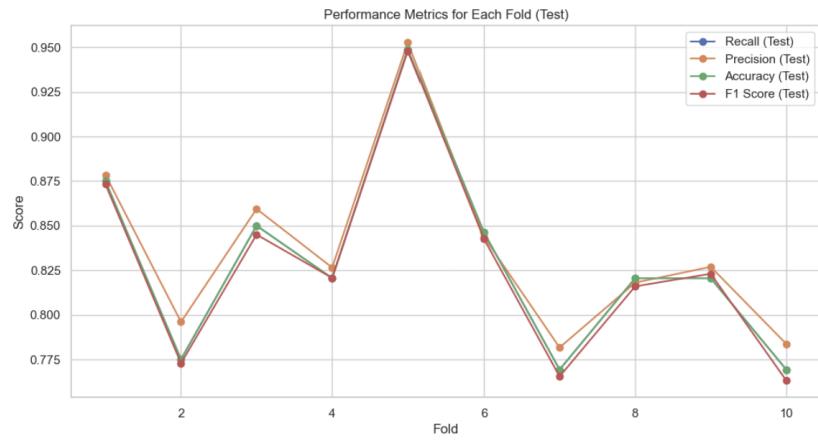


Figura 141: Model SVM final

En aquest cas podem observar com el recall es manté dins d'un bon rang, entre 0.8 i 0.95. A més a més, la precisió mostra una estabilitat semblant al recall, amb valors. La resta de mètriques, mostren una tendència similar a aquestes dues, demostrant un bon rendiment.

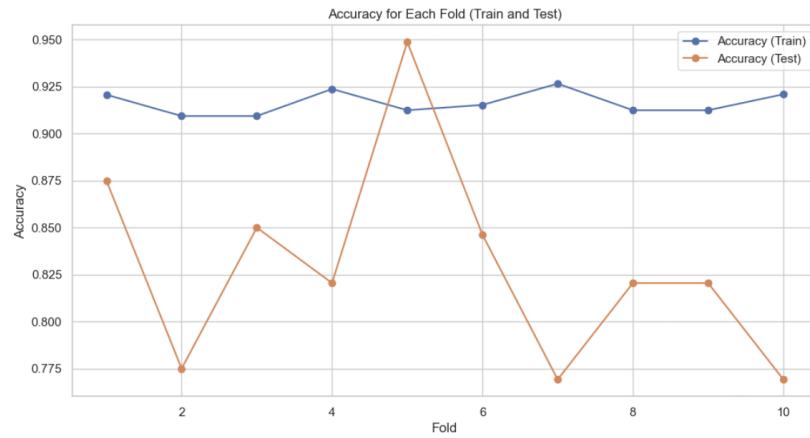


Figura 142: Model SVM final

En la segona imatge podem veure com l'exactitud d'entrenament és consistentment alta, prop de 0.925, indicant que el model aprèn eficaçment dels dades d'entrenament. En el cas del test, veiem com aquest valor disminueix però malgrat tenir una certa variabilitat, els punts més baixos no cauen molt per sota del 0.8 i el model recupera ràpidament, el que demostra tenir una bona capacitat de recuperació davant de variacions en les dades.

Finalment, analitzant la diferència entre l'exactitud mitjana d'entrenament i de test podem dir que aquesta és relativament petita, el que és una bona indicació de la capacitat del model per generalitzar a noves dades sense caure en el sobreajustament.

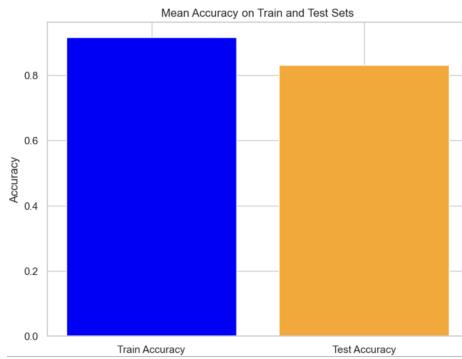


Figura 143: Model SVM final

## 6.4 EBM final

Podem observar que el millor preprocessament en aquest cas és no tractar els outliers, eliminar les missings de les variables drug i després aplicar KNN per imputar les variables numèriques. Per balancejar, utilitzar smotenc i normalitzar amb MinMax.

Així doncs, els millors paràmetres són un learning rate de 0.01, un màxim de 100 bins, un màxim de 5 iteracions i un màxim de 100 rondes.

Procedim doncs, a analitzar els resultats obtinguts.

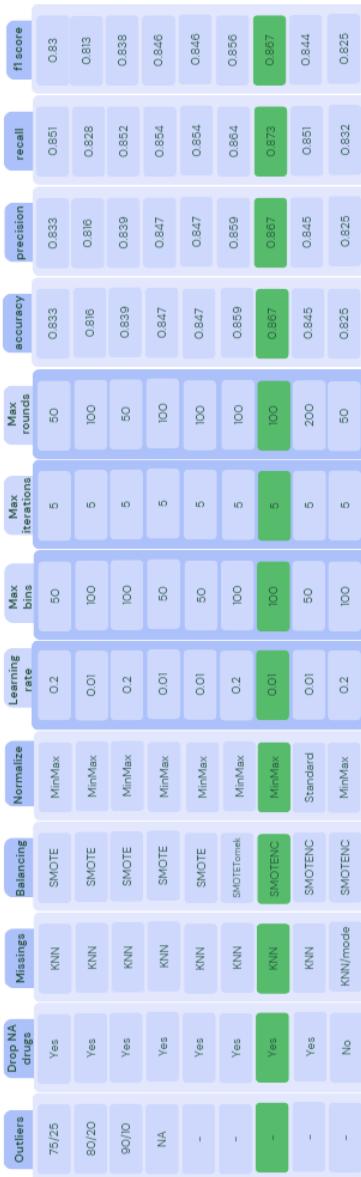


Figura 144: Proves EBM

Observant la imatge de la següent pàgina, podem observar com en aquest cas totes les mètriques segueixen una tendència similar. A més a més, el rendiment es troba entre el 0.76 i el 0.93. Podem observar una variabilitat especialment notable en els folds 7 i 8 el que ens pot indicar que el model és sensible a un cert tipus de dades contingut en aquests folds. El fet que les línies de les mètriques estiguin tan juntes, proposa que el model està ben balancejat entre precision i recall.

En el cas de la figura podem observar com l'accuracy de l'entrenament es manté constant en valors alts. No obstant, l'accuracy del test, presenta més variabilitat i se situa en valors menors. En aquest cas podriem tenir un cas d'overfitting.

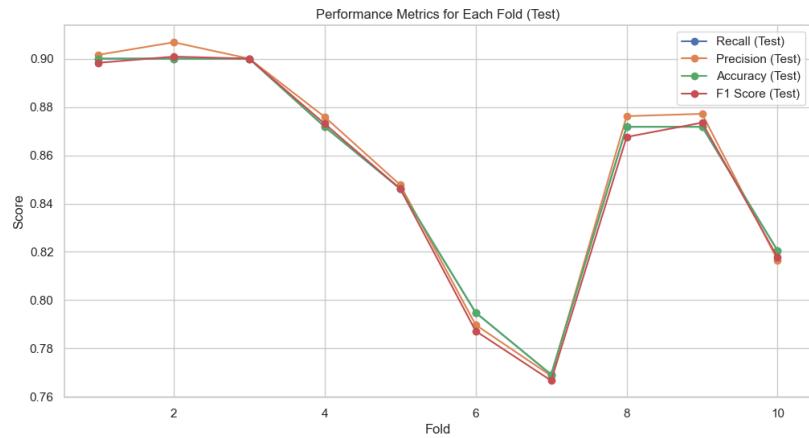


Figura 145: Model EBM final

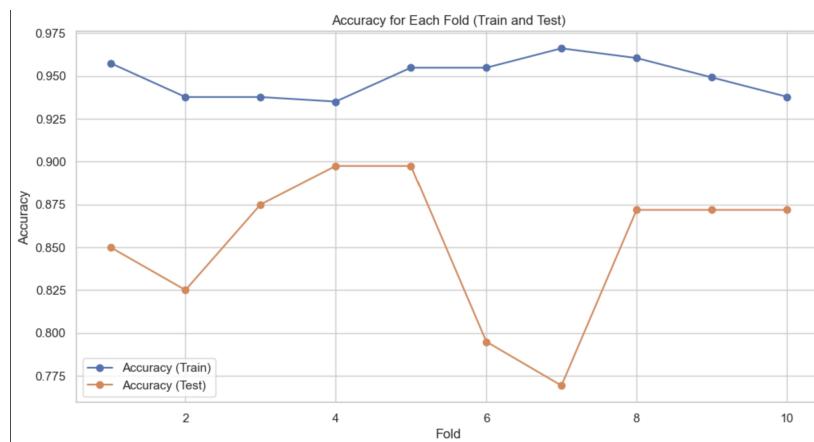


Figura 146: Model EBM final

Finalment, observant l'histograma podem observar com l'accuracy del train és superior a la del test, no obstant, aquesta diferència és molt petita el que ens indica que el model generalitza correctament.

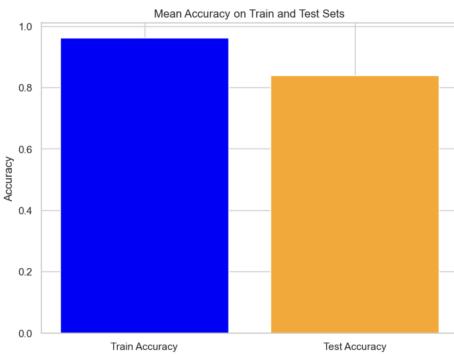


Figura 147: Model EBM final

Per poder determinar quines variables són les més important en l'algorisme, vaig generar les gràfiques següents:

En l'eix X es representa cada valor de la variable age, en aquest cas, entre 0 i 1 ja que està



Figura 148: Variable age en el EBM

normalitzada. Per altra banda, l'eix Y indica l'score o els punts de contribució al pronòstic que la característica age aporta al model. Així doncs, un valor alt en l'eix Y indica una major contribució positiva a la predicción de la variable objectiu. En aquest cas podem veure la contribució en cada variable.

A la gràfica de sota, podem veure la gràfica de distribució de la variable. Així doncs, podem veure com la variable age té un efecte significatiu en la predicción de la variable objectiu. En aquest cas,



Figura 149: Variable cholesterol en el EBM

també he detectat que la variable cholesterol és important en la predicción. Podem veure com sobretot és important per predir la classe 2.

Així doncs, després d'analitzar aquestes gràfiques per cada variable, vaig poder determinar que les variables més significatives i importants per la predicción de la variable objectiu en l'algorisme EBM són: la variable age, la variable cholesterol, la variable ascites, la variable edema, la variable Alk\_Phosphatase, la variable prothrombin, la variable platelets, la variable stage i la variable N\_Years.

A continuació es poden trobar les gràfiques que ho demostren.



Figura 150: Variable ascites en el EBM



Figura 151: Variable edema en el EBM



Figura 152: Variable Alk\_Phosphatase en el EBM



Figura 153: Variable Prothrombin en el EBM

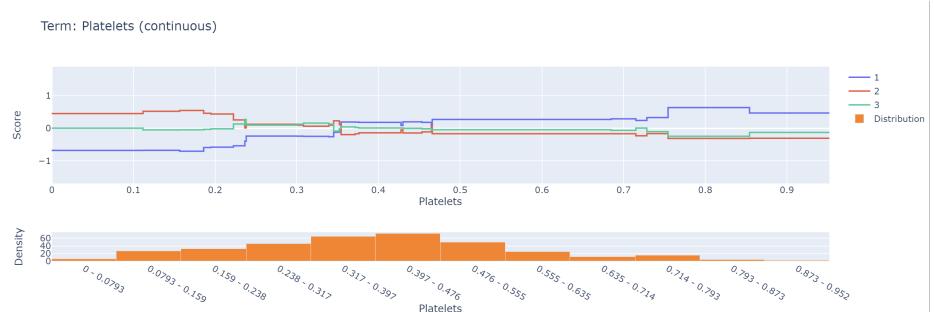


Figura 154: Variable Platelets en el EBM

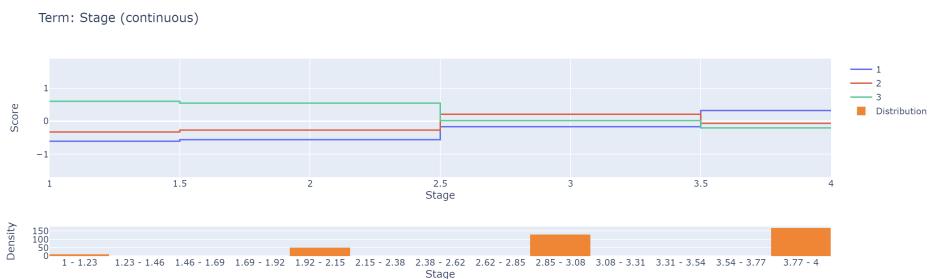


Figura 155: Variable Stage en el EBM



Figura 156: Variable N\_Years en el EBM

Així doncs, un cop vaig escollir les millors models per cada algorisme, va ser el pas d'escollir entre ells un model final.

Després d'analitzar les gràfiques de cada model, vaig decidir quedar-me amb els models SVM i EBM.

En el cas del KNN, el fet que l'accuracy de l'entrenament fos 1, indica que el model podria estar generalitzant les dades. A més a més, analitzant les gràfiques de les diferents mètriques es pot veure que hi ha folds en que el model té unes mètriques molt menors a la resta de folds, el que pot indicar que no és un model constant sinó que el seu rendiment es basa en el conjunt de dades que li dones per predir, i per tant, no pot tenir un rendiment constant.

Per altra banda, en el cas de l'arbre de decisió, en aquest podem trobar una diferència notable entre l'accuracy de l'entrenament i l'accuracy del test, el que podria indicar que l'algorisme pateix de sobreajustament a les dades d'entrenament.

Per altra banda, en el cas del SVM, podem adonar-nos que també té certa variabilitat entre els folds, no obstant això, aquesta variabilitat no varia tan o no disminueix tan com en el cas del KNN. A més a més, podem adonar-nos que també hi ha variabilitat en el conjunt d'entrenament que no és sempre 1, el que podria allunyar la idea de patir sobreajustament a les dades d'entrnamet. De fet, observant l'histograma, es pot observar com la diferència entre les accuràcies de l'entrenament i la del test, no és tan notable.

En el cas del EBM, podem veure com les mètriques tampoc són constants i que poden dependre de les dades d'entrada. No obstant això, observant l'histograma on es compara l'accuracy de l'entrenament i del test, es pot observar com la diferència és poc significativa, i el nivell d'accuracy del test és alta.

Així docns, un cop decidits que els models serien el model EBM i el model SVM, vaig decidir aplicar-los a les dades del test per observar com es comportaven i qui dels dos era preferible.

SVM:

Classification Report on Test Set:				
	precision	recall	f1-score	support
1	0.22	0.33	0.27	6
2	0.58	0.75	0.65	20
3	0.79	0.59	0.68	37
accuracy			0.62	63
macro avg	0.53	0.56	0.53	63
weighted avg	0.67	0.62	0.63	63

Figura 157: Test SVM

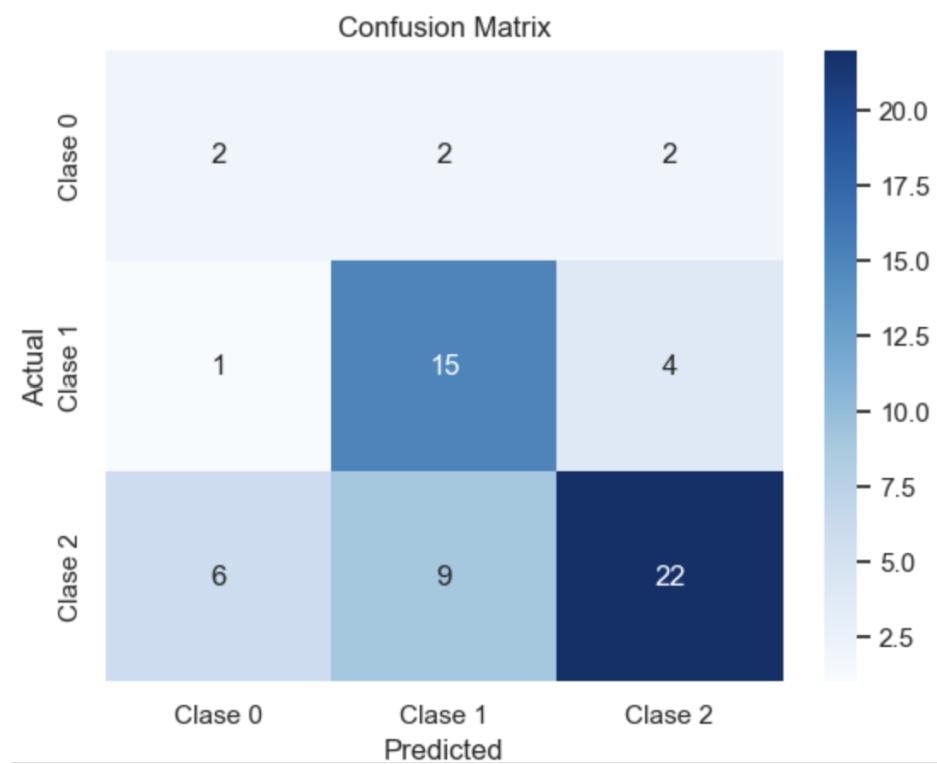


Figura 158: Test SVM

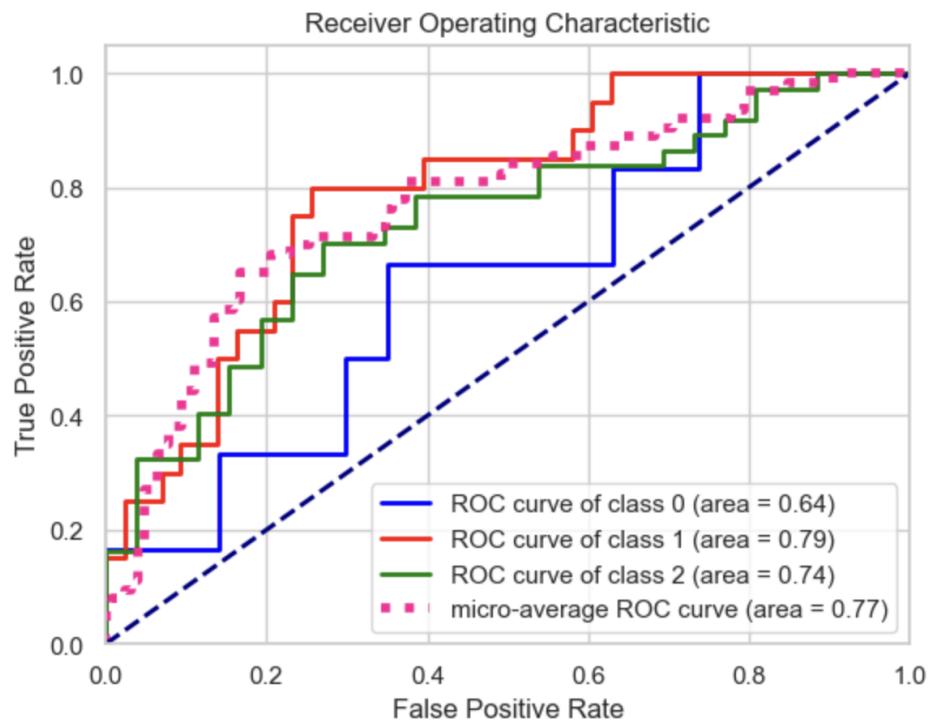


Figura 159: Test SVM

**EBM:**

Classification Report on Test Set:					
	precision	recall	f1-score	support	
1	0.00	0.00	0.00	6	
2	0.64	0.80	0.71	20	
3	0.80	0.76	0.78	37	
accuracy			0.70	63	
macro avg	0.48	0.52	0.50	63	
weighted avg	0.67	0.70	0.68	63	

Figura 160: Variable N\_Years en el EBM

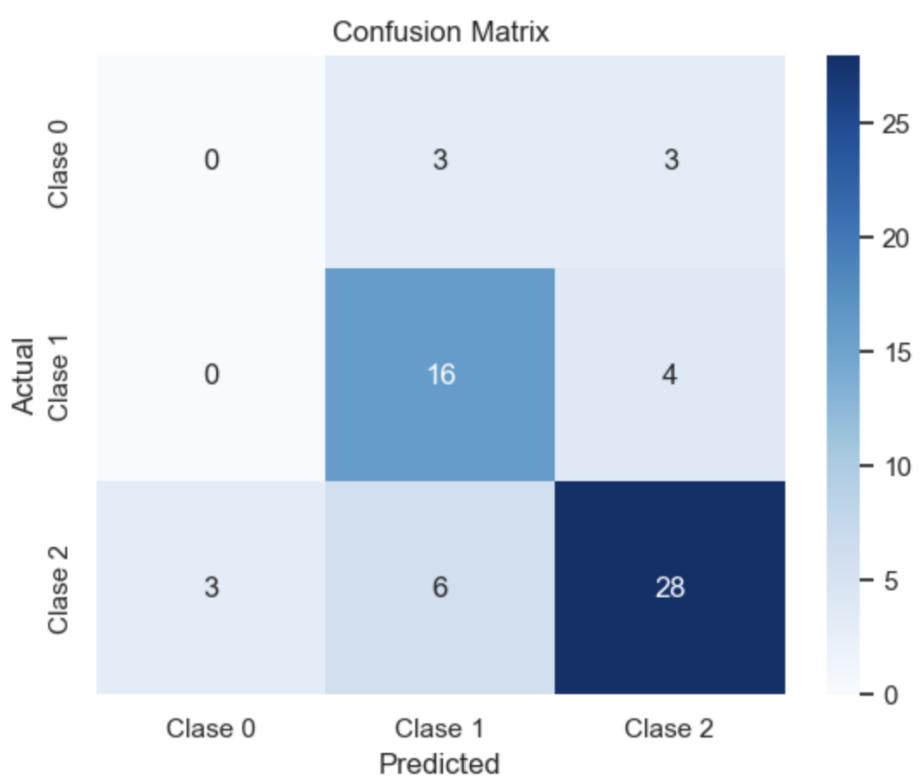


Figura 161: Variable N\_Years en el EBM

Per poder escollir el model més indicat, analitzaré els dos models a la vegada.

Amb l'informe de classificació i la matriu de confusió de cada model, podem observar que la classe 1, és a dir, la categoria 'CL', obté un recall i una precisió de 0 en el model EBM, el que indica que el model no ha pogut classificar bé cap mostra d'aquesta classe.

En el cas del model SVM, podem observar com la predicció d'aquest model en aquesta classe, millora respecte l'altre model, amb una precisió de 0.22 i un recall de 0.33, no obstant, aquests valors són baixos.

Per la classe 2, corresponent a 'D', podem observar com té una precisió decent, de 0.64 i un recall alt de 0.8, el que indica que el model és bastant bo en la detecció d'aquesta classe, però que encara confon algunes mostres d'altres classes com a 2.

En el cas de la classe 2 per l'algorisme SVM, aquest té una precisió més baixa de 0.58 en comparació amb el model EBM, però el recall és gairebé igual (0.75).

I finalment, en el cas de la classe 3, 'C', l'algorisme té la millor precisió (0.8) i un bon recall de 0.76, el que indica que el model realitza correctament la classificació d'aquesta classe.

Pel cas de l'SVM, la precisió en la classe 3 és similar al primer model, amb un 0.79, però el re-

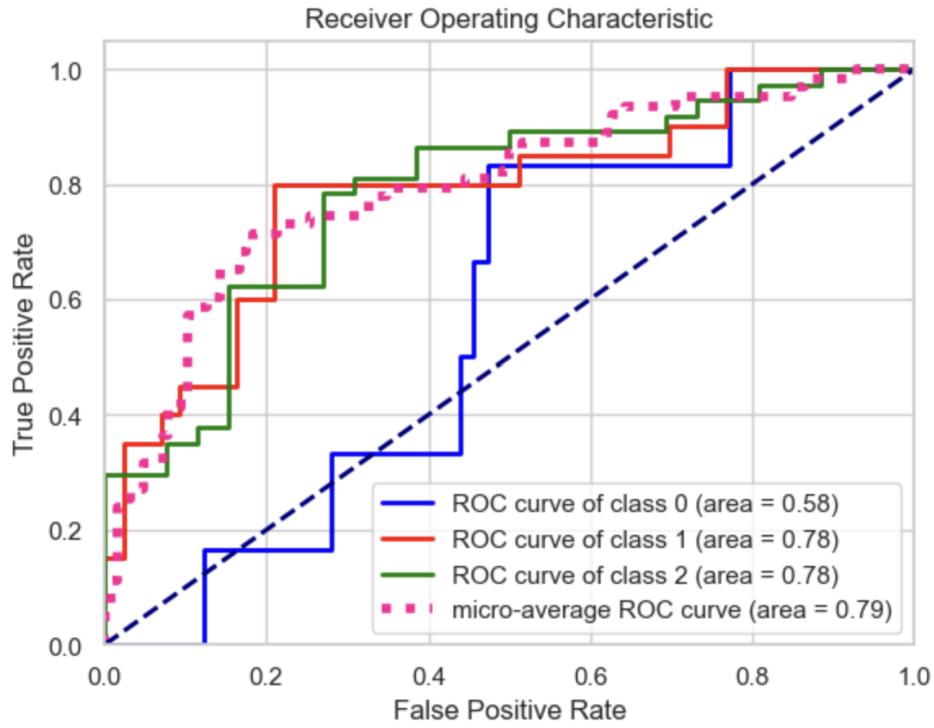


Figura 162: Variable N\_Years en el EBM

call és més baix (0.59).

Analitzant les exactituds generalis, podem dir que l'exactitud de l'EBM és força alta (0.7), però el poc rendiment a la classe 1 fa que els valors macro i ponderats mitjans de precisió i recall siguin baixos.

Per altra banda, en el cas de SVM, l'exactitud general és menor a la de l'EBM (0.62) i els valors macro i ponderats són també més baixos que en EBM, però més equilibrats entre classes.

No obstant això, el fet que en els dos casos l'accuracy disminueixi considerablement respecte l'accuracy del test, pot indicar que els dos pateixen d'overfitting. Una consideració per millorar aquest overfitting podria ser utilitzar més dades per entrenar el model.

Analitzant les curbes roc, podem veure que l'àrea AUC és major en el cas de l'algorisme EBM, tot i que per aquest cas, la curva ROC de la classe 0 presenta un rendiment molt baix, a més a més, la diferència entre les àrees dels diferents models no és molt notable. Pel model EBM, podem veure un comportament similar en les curbes ROC de les classes 1 i 2.

En la curva ROC de SVM, podem observar com la curva de la classe 0 també té una àrea menor respecte a les altres classes i que la classe 1 i 2 es comporten similar, no obstant, l'àrea de la classe 0 és superior a l'àrea de la classe 0 però en el cas de EBM.

En aquest cas, la decisió de quin model és millor, la deixaria en mans d'un professional, exposant-li els

casos. L'algorisme SVM presenta un rendiment menor en les classes 'C' i 'D' però prediu mínimament bé els casos 'CL'. En canvi, l'algorisme EBM presenta un rendiment major en les classes 'C' i 'D' però no predeix bé la classe 'CL'. En funció de les classes que prioritzen i del tipus d'error que prefereixin, esculliria un model o un altre. Si és crític no perdre's cap cas de la classe 1, SVM pot ser preferible, si és més important ser molt segur a les prediccions de la classe 3 (C), és preferible l'EBM.

Si es necessita un model que funcioni millor entre els tres casos, SVM és preferible, en canvi si es necessita un model amb major exactitud, és preferible l'EBM.

En aquest cas, per escollir el model final em vaig basar en les dades inicials. Tal com podem veure, la classe minoritària és CL, el que pot indicar que el model potser no ha trobat prous casos per entrenar correctament o per comprovar la seva precisió i el seu recall en aquesta classe. A més a més, el fet que hi hagi tan pocs individus de la classe, pot voler indicar que no és una classe majoritària i tan important com les altres. I finalment, vaig fer la reflexió que el model es pot simplificar de forma que determini si o bé sobreviu o bé mor, que de fet, aquest és l'objectiu del problema. Per tant, un cop considerat això, vaig decidir que era més important l'exactitud total davant l'exactitud entre classes i que per tant, em quedava com a model final el model EBM, malgrat les dificultats d'aquest per predir la classe 'CL'.

## 7 Model Card: Predictor de supervivència en pacients amb cirrosi hepàtic

### Details del Model

#### Model

Aquest model prediu si un pacient amb cirrosis hepàtic sobreuirà, sobreuirà amb trasplantament de fetge o morirà, a través de diferent informació mèdica. El model s'ha entrenat amb l'algorisme Explainable Boosting Machine. Els hiperparàmetres són un learning rate de 0.01, un màxim de 100 bins, un màxim de 5 iteracions i un màxim de 100 rondes. Aquest model combina models lineals amb arbres de decisió.

#### Versió

- Nom: CirrosisSurvivalPredictor
- Data: 28/12/2023
- Tipus: Classificació

#### Propietaris:

- Nom: Núria Llopert Fernàndez (estudiant IAA a la UPC), nuria.llopert@estudiantant.upc.edu

#### Referències:

- Conjunt de dades: BaseDades
- Scikit-Learn: Scikit-Learn

## **Consideracions**

### **Usuaris esperats**

- Professor IAA
- Estudiant IAA

### **Ús**

- La intenció d'aquest model és que com a estudiant de IAA, aprengui a realitzar un model i analitzar-ne els resultats. Aquest model no té cap intenció en ser usat per analitzar cap pacient real. Només ha de ser servit per analitzar i puntuar el meu treball.

### **Limitacions**

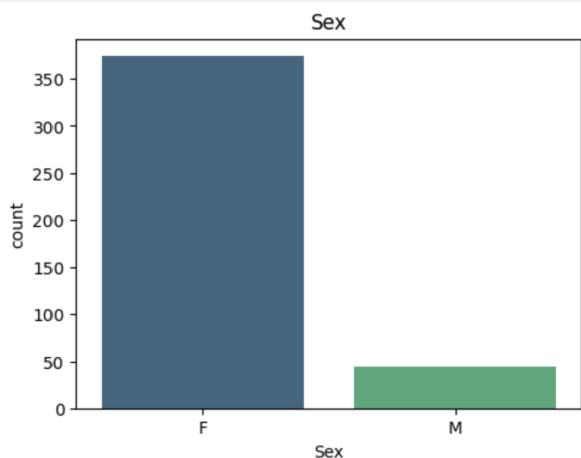
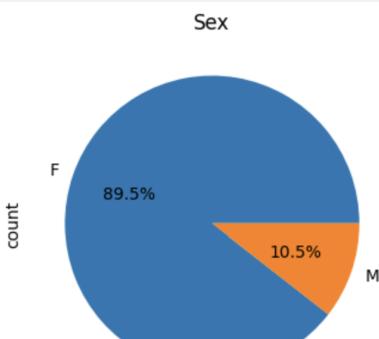
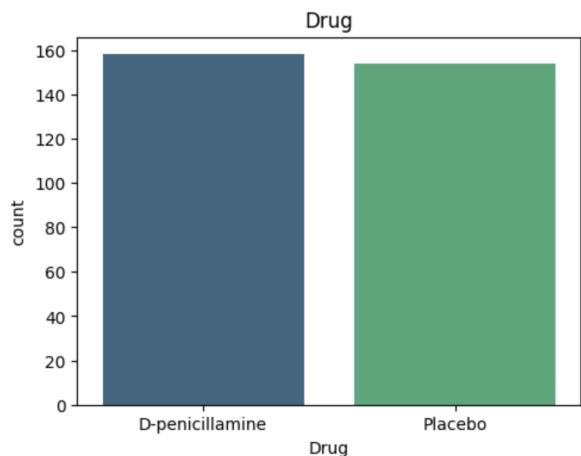
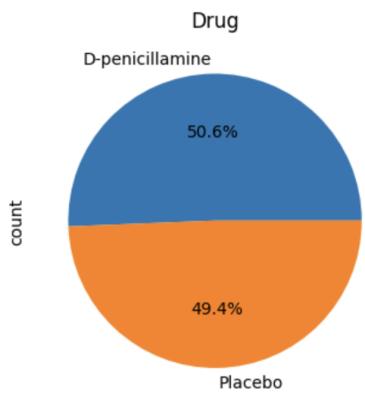
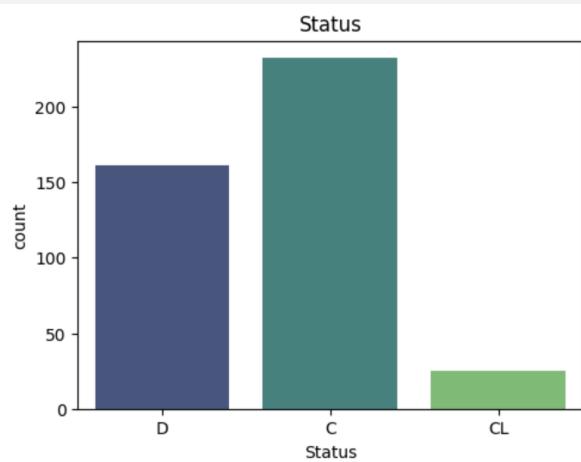
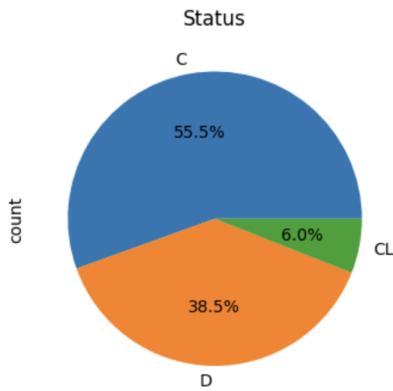
- El model pot no ser capaç o predir malament les classes minoritàries, en el nostre cas la classe 'CL'.
- El model pot patir de sobregeneralització a les dades d'entrenament i pot fer prediccions incorrectes amb dades noves, especialment si aquestes són molt diferents de les dades utilitzades per entrenar.

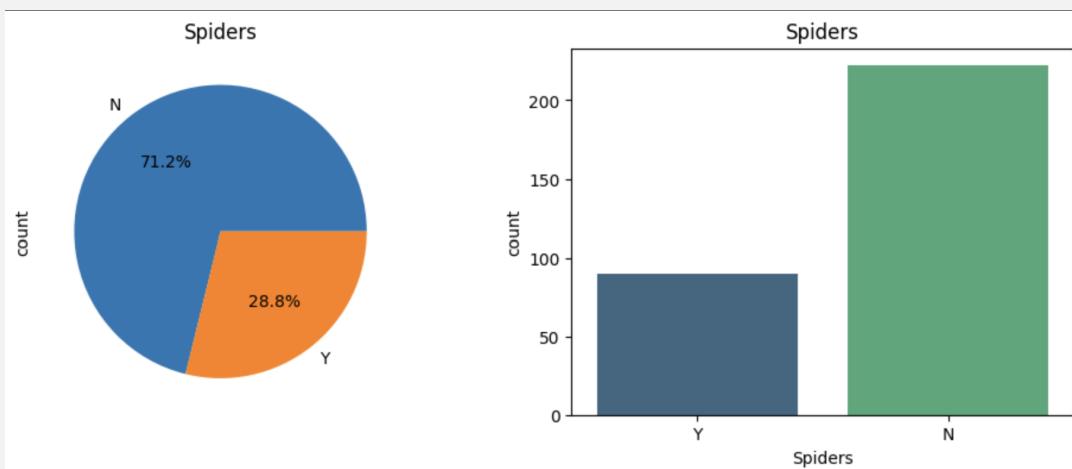
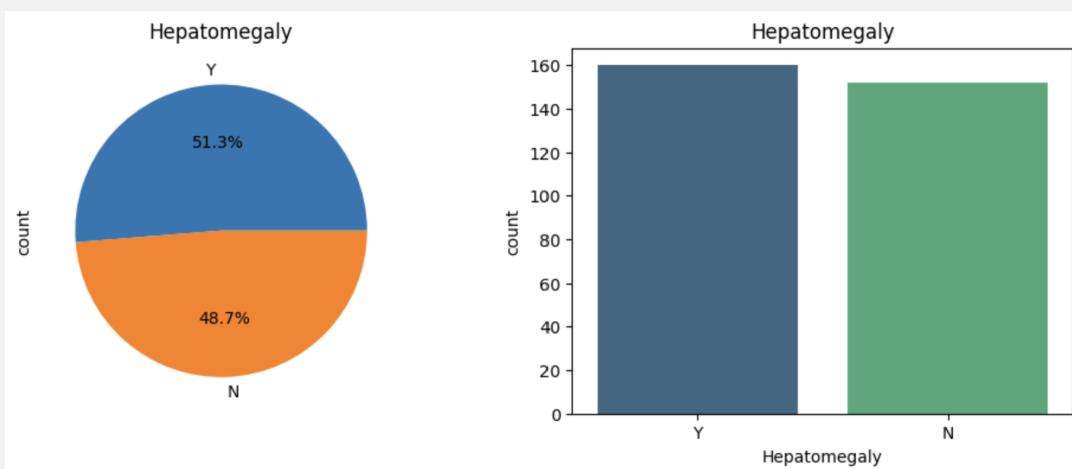
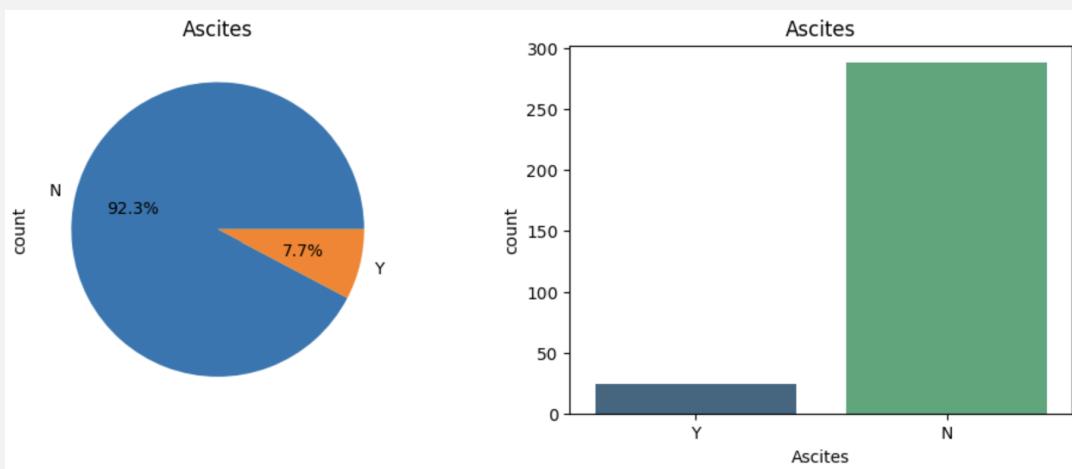
### **Consideracions ètiques**

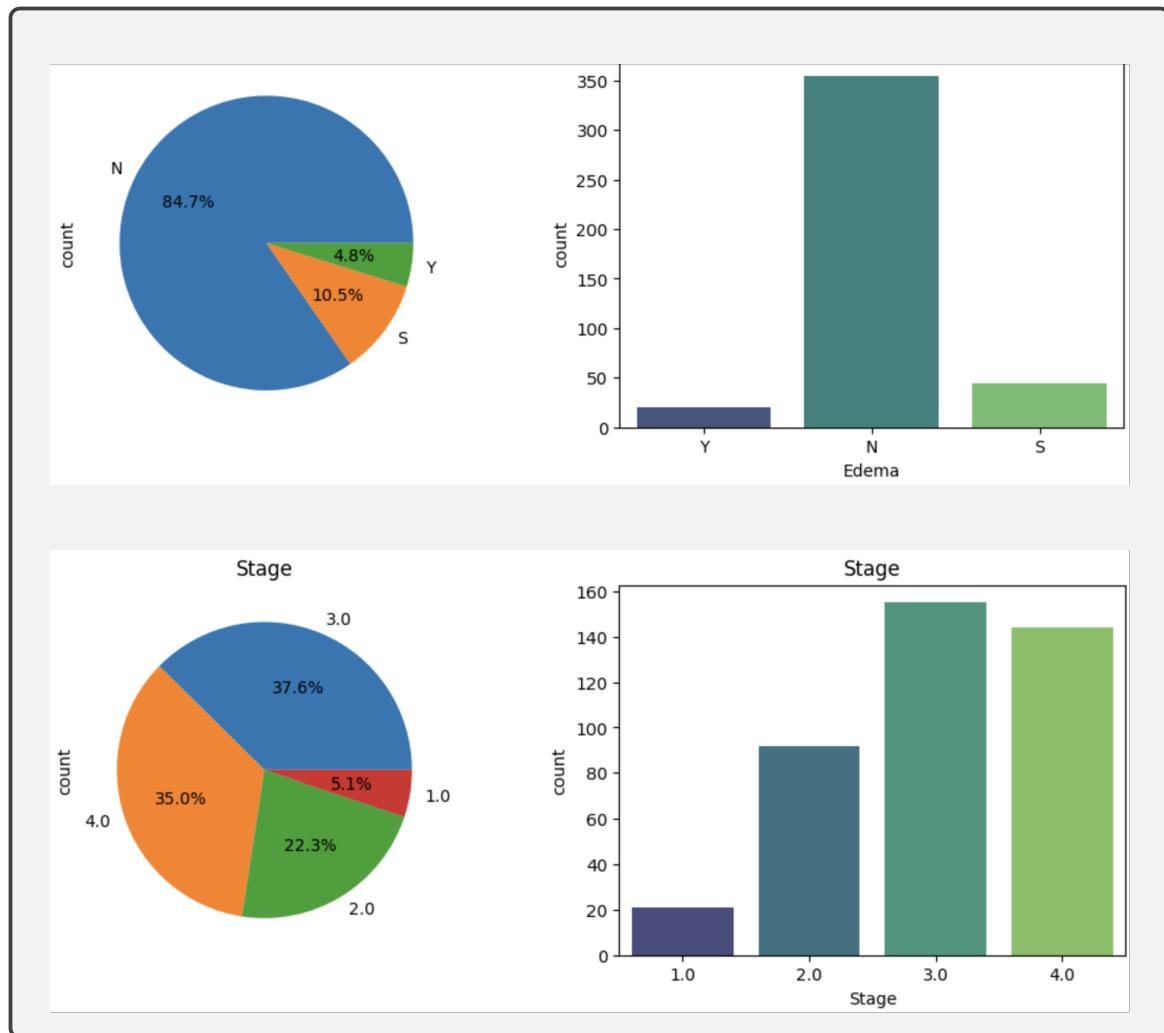
- Si el model es volgués portar en un hospital, caldria que el pacient conegués el model i els riscos que presenta.

## Datasets

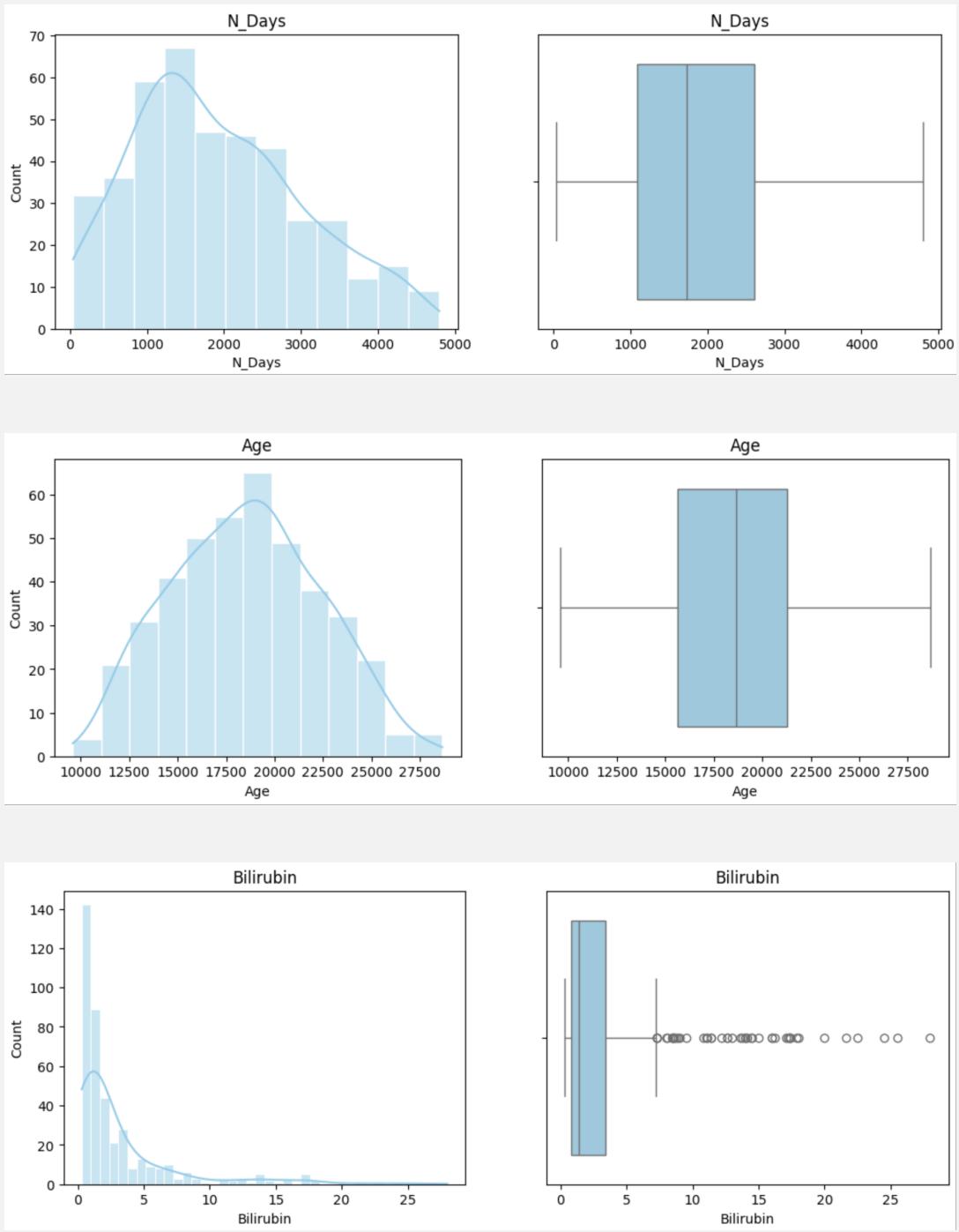
### Variables categòriques

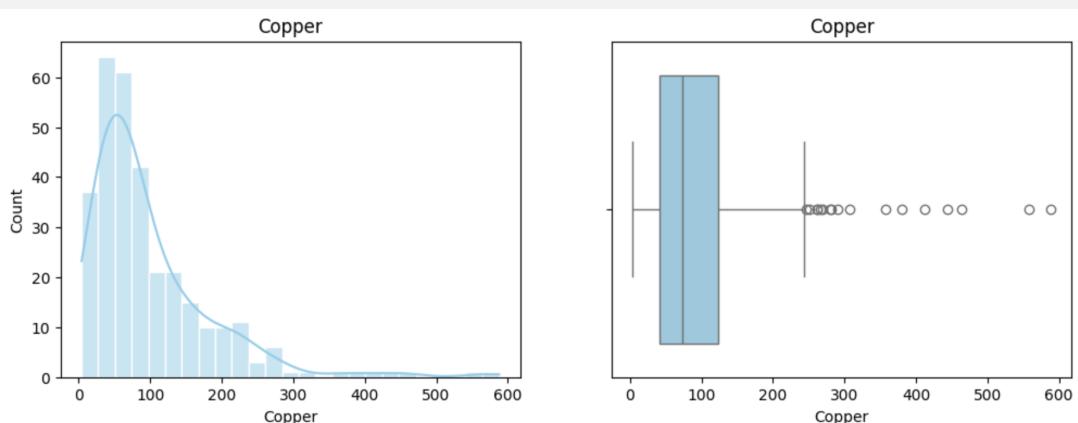
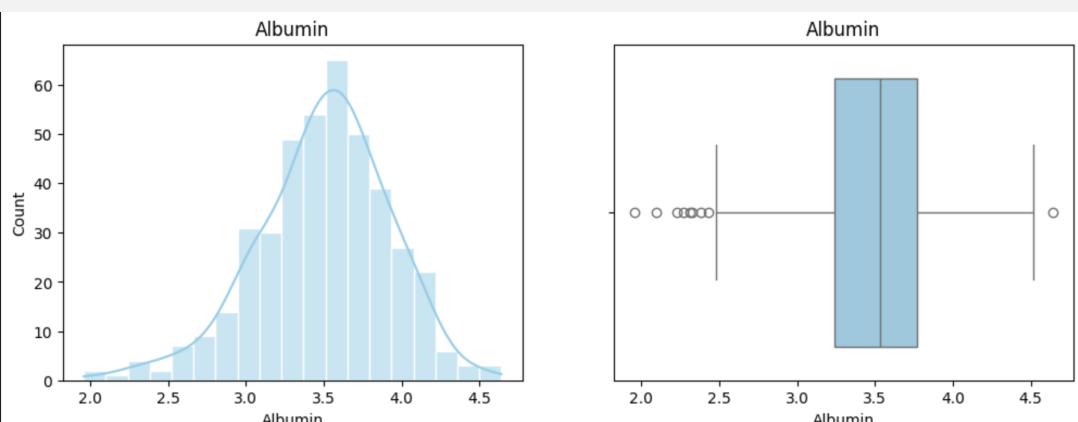
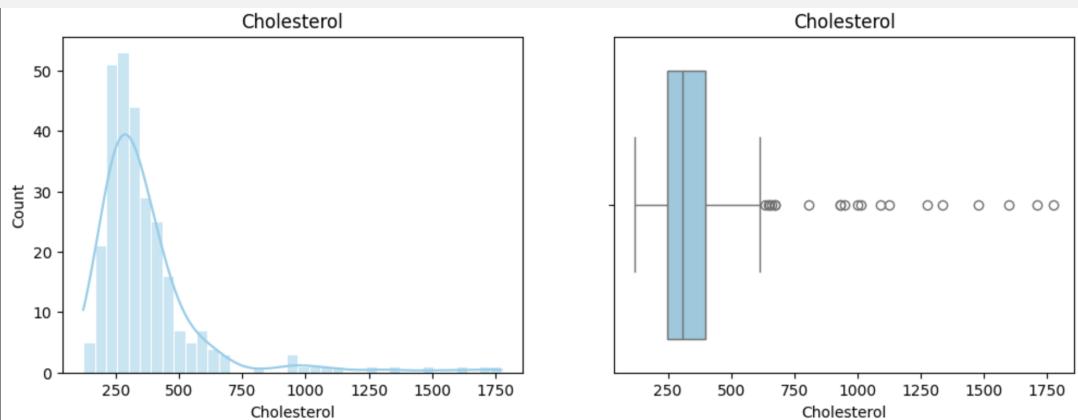


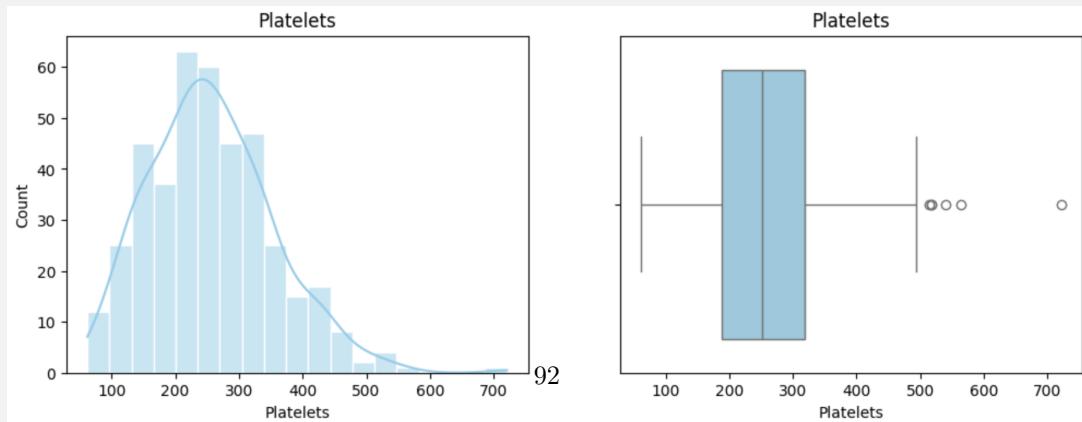
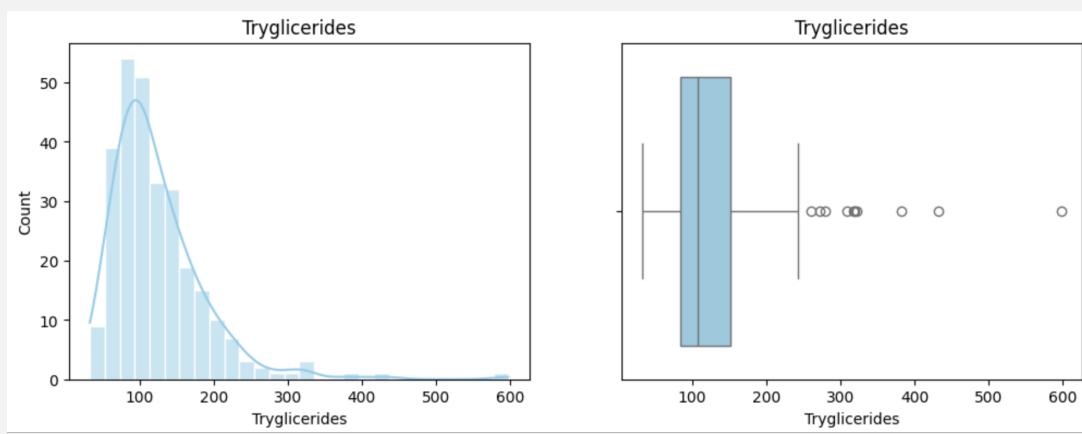
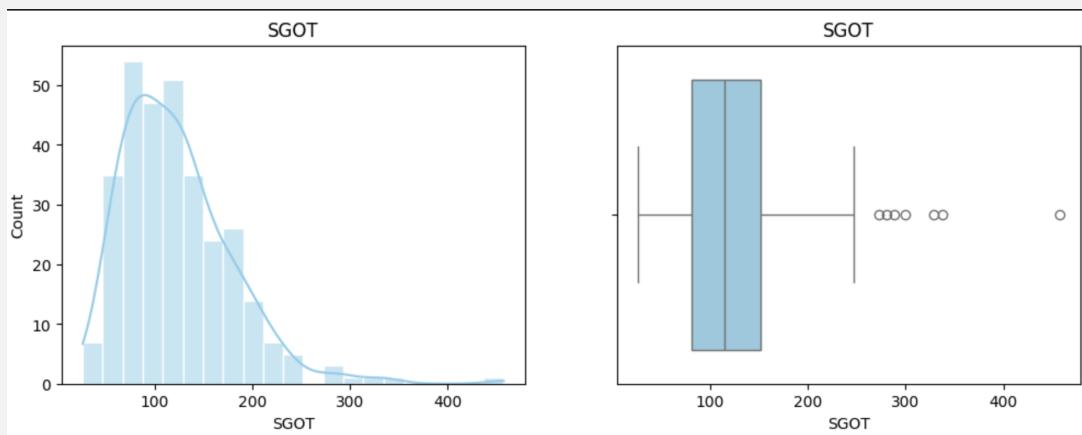
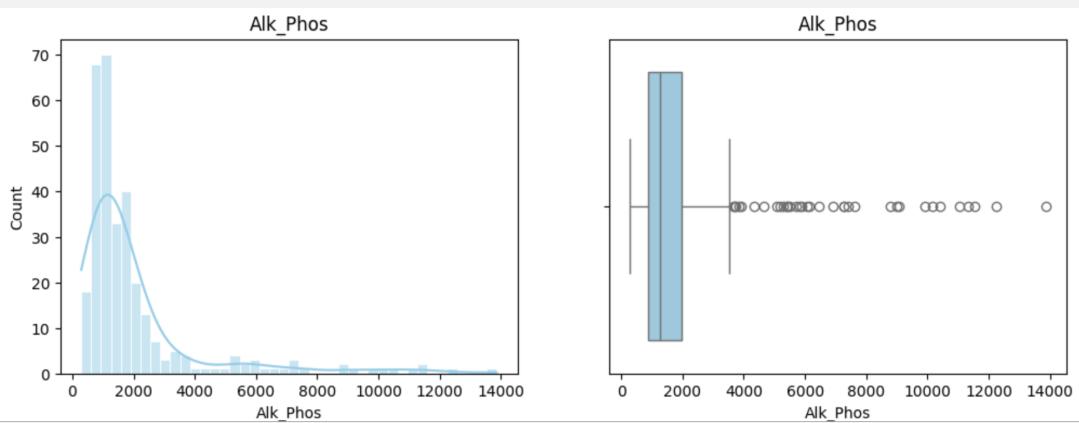


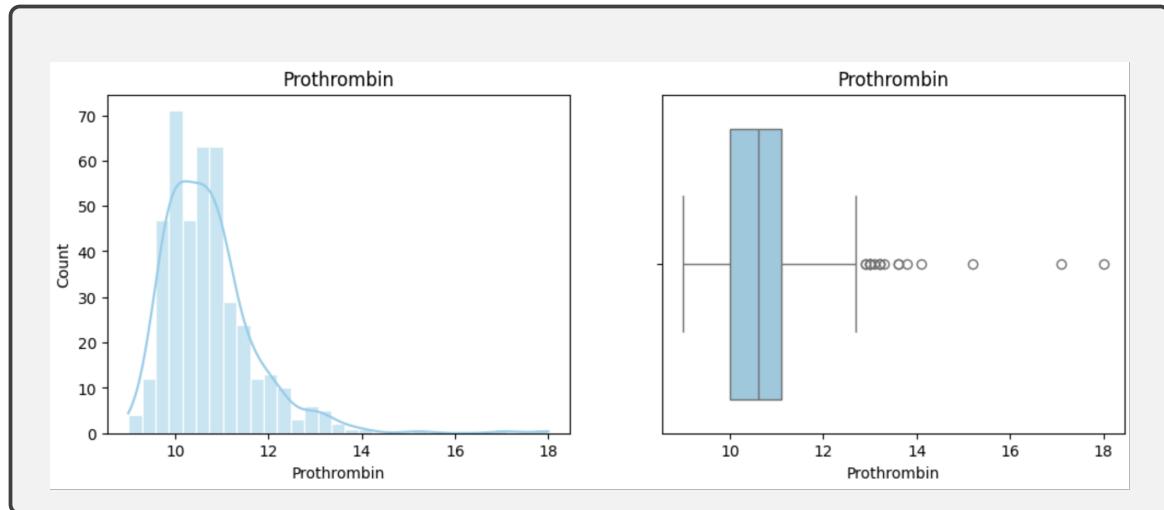


## Variables numériques



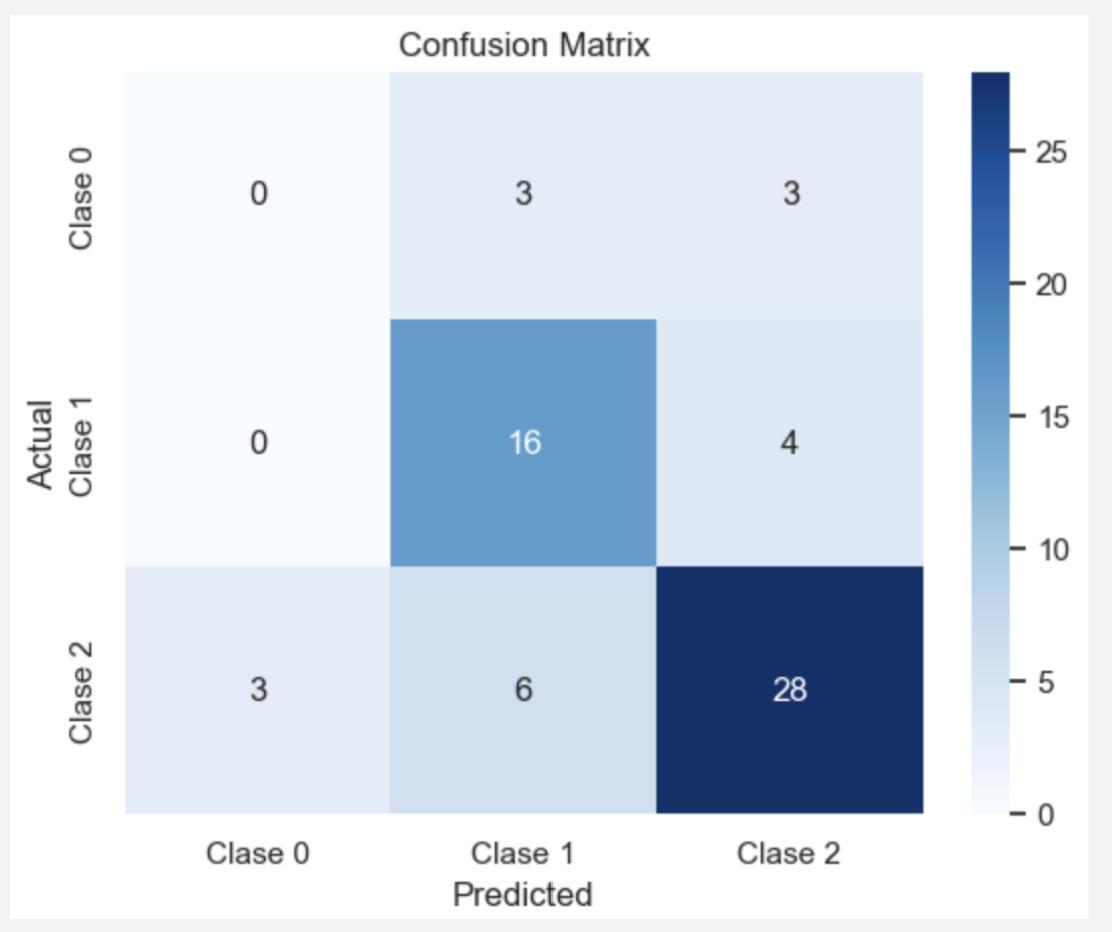




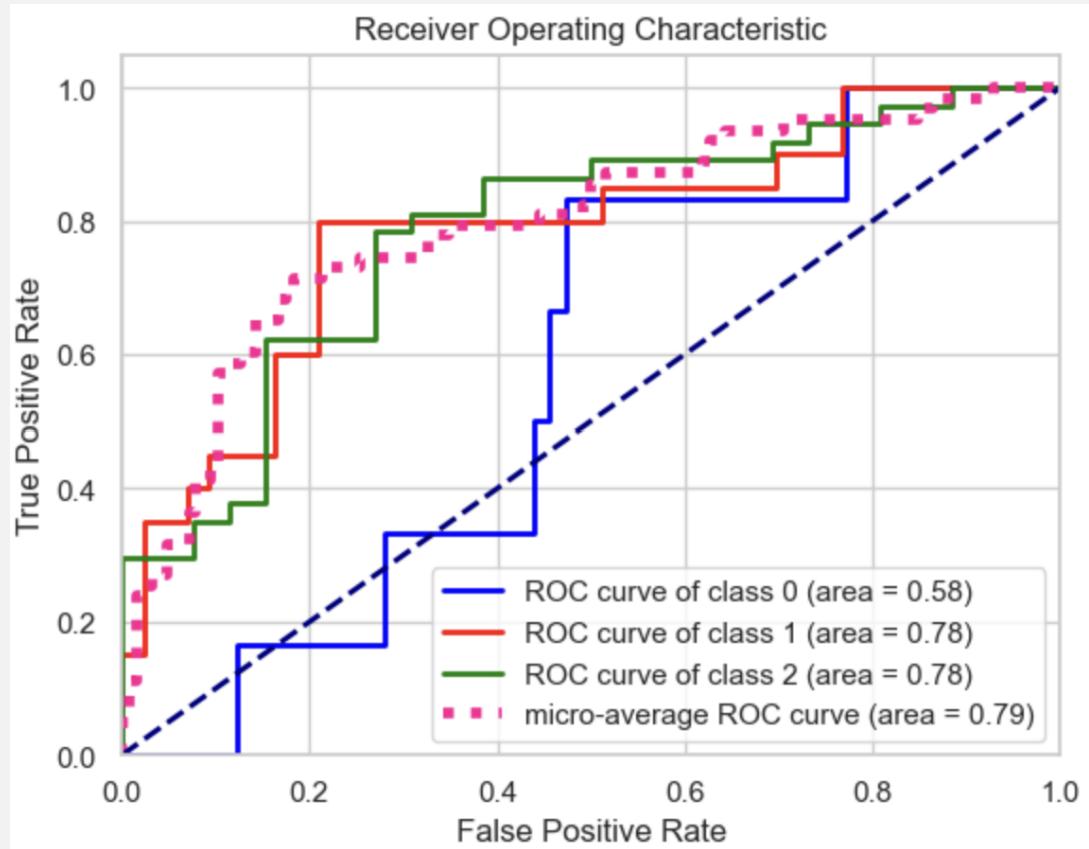


## Anàlisi Quantitativa

### Matriu confusió



Curva ROC:



## 8 KMeans i Hierarchical Clustering

Finalment, vaig realitzar el KMeans i el Hierarchical Clustering per analitzar si podem identificar clústers útils per la nostra tasca.

Després d'aplicar el Kmeans amb K=3, ja que volia provar de predir les tres classes de la variable objectiu, vaig obtenir el següent: Tal com podem veure a la imatge, no aconsegueix classificar els

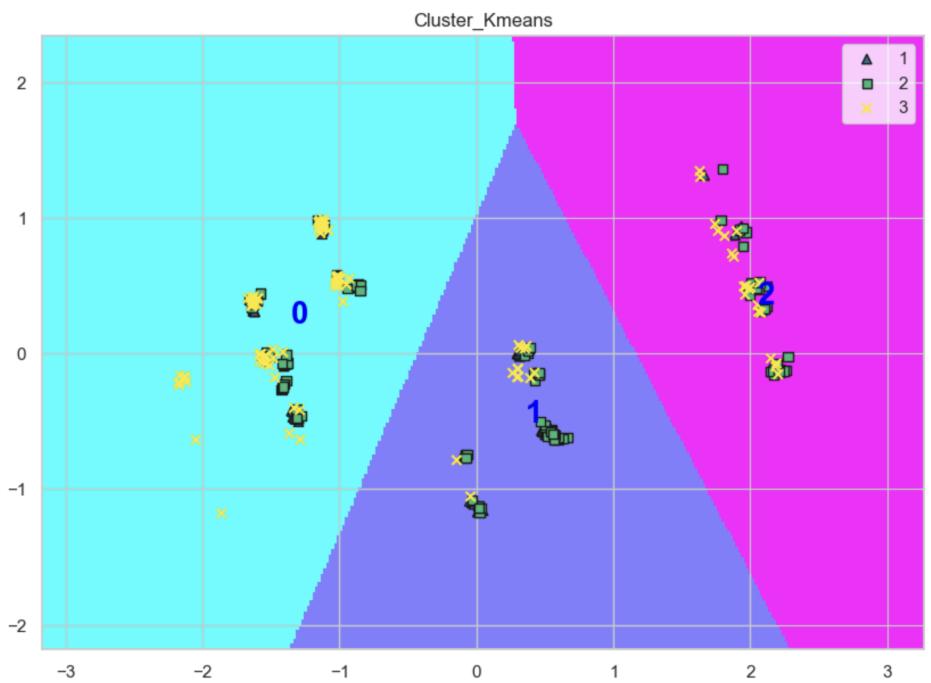


Figura 163: KMeans

clústers segons la variable objectiu, per tant, el KMeans, no seria útil per la nostra tasca.

No obstant això, vaig voler mirar les diferències significatives per cada variable en cada clúster.

A la imatge següent podem veure les diferències significatives entre el clúster 0 i el clúster 1. Podem observar com les variables que més diferències mostren són l'edema l'alk\_phos i els triglicèrids.

Per altra banda, trobem diferències moderades en la bilirubina i el copper.

La resta de variables mostren poca o cap diferència significativa entre aquests clústers.

En el cas del clúster 0 i el clúster 2, podem observar com la majoria de la diferència es troba en Alk\_Phosphat i els triglicèrids mentre que el colesterol i l'edema mostren algunes diferències però que són baixes. La resta, no mostra diferències significatives.

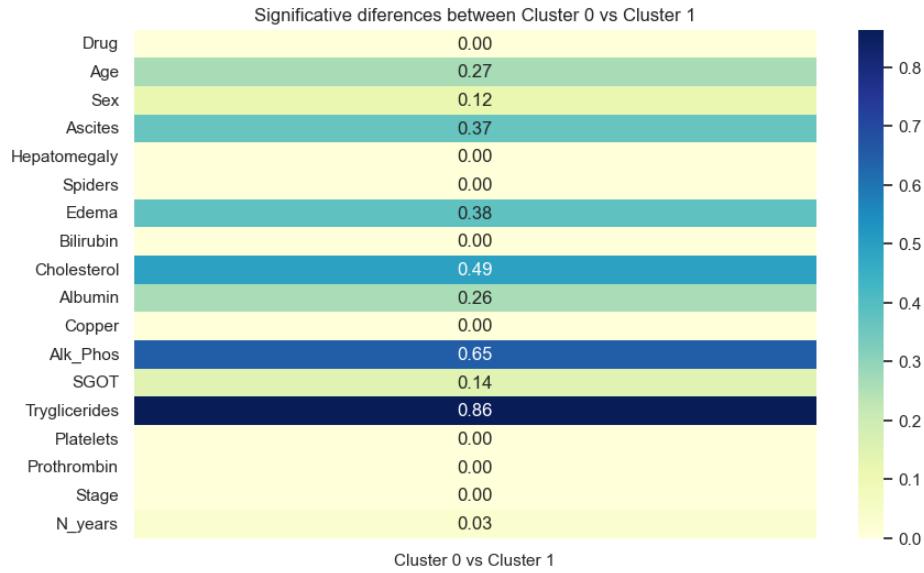


Figura 164: Diferències significatives clúster 0 i 1

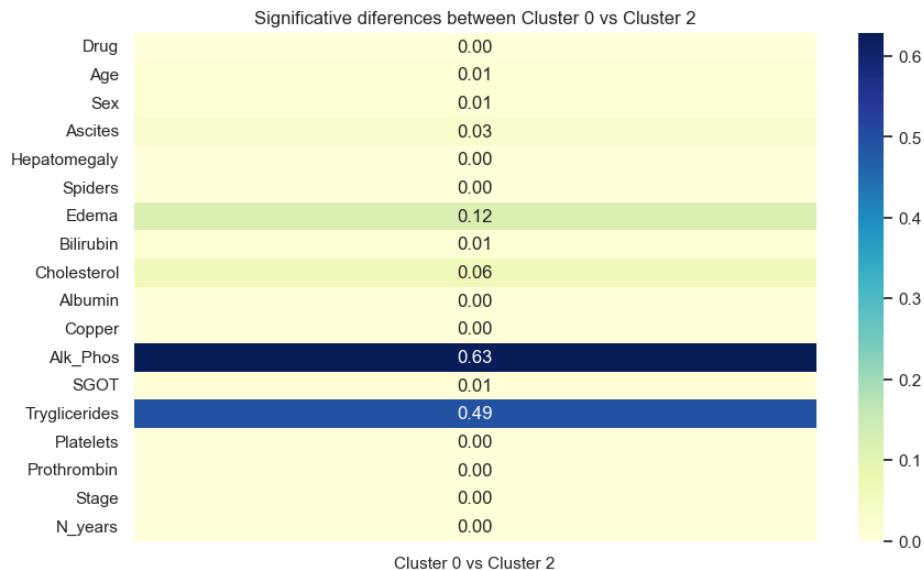


Figura 165: Diferències significatives clúster 0 i 2

En el cas dels clústers 1 i 2, podem notar que la diferència més gran es troba en els triglicèrids, Alk\_Phos i albumina, mentre que l'edat i el sexe mostren diferències significatives.

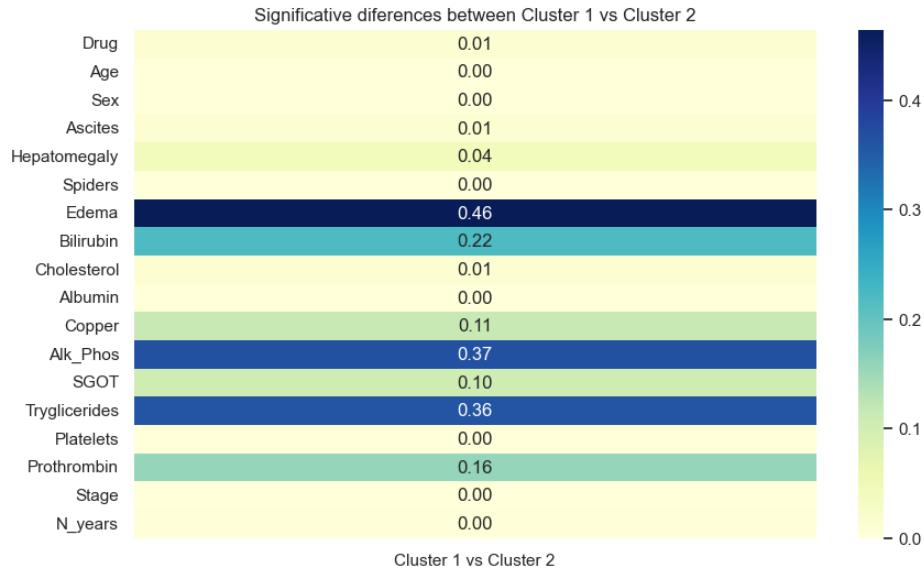


Figura 166: Diferències significatives clúster 1 i 2

Així doncs, podem conoure que els grups creats semblen diferenciar-se clarament segons els nivells d'Alk\_Phosphatase i triglicèrids. També hi ha algunes diferències en l'edema.

En veure que el KMeans no funcionava, vaig voler aplicar el hierarchical clustering per veure si funcionava millor.

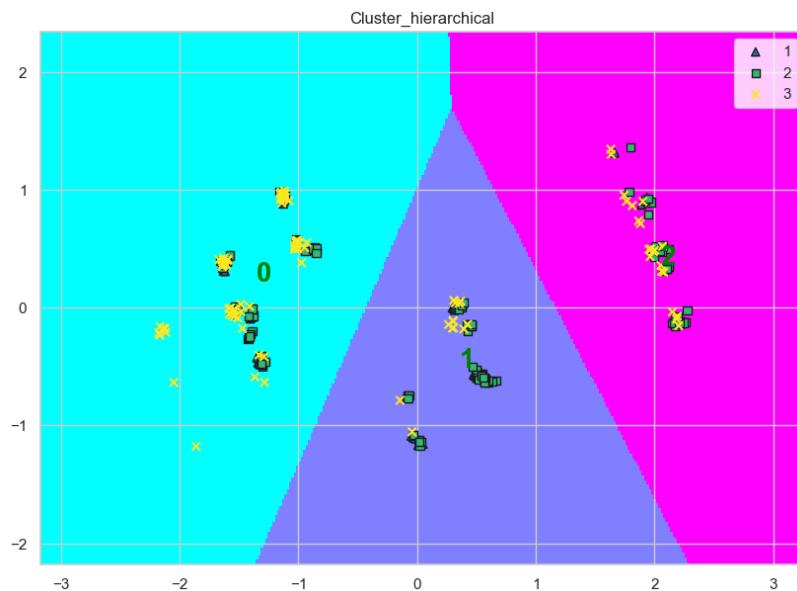


Figura 167: Hierarchical Clustering

Vaig observar que em creava els mateixos clústers que en el cas del KMeans. En aquest cas, vaig decidir mostrar l'arbre que ens crea, tal com es pot veure a la figura 168. Seguidament, tal com es pot veure a la figura 169, vaig intentar tallar-los a un punt lògic per veure la quantitat de clústers obtinguts i si aquest s'adaptava als que havia imposat, és a dir 3. Tal com es pot veure a l'arbre, no hi ha cap punt en el qual tallar en els quals hi hagi tres grups, com a molt, podem trobar un amb dos que podria ser els que sobreviuen i els que moren. Per tant, hierarchical clustering tampoc ens serveix per la nostra labor, caldria estudiar si aquest serveix per predir tan sols dues classes, si una persona morirà o no.

## 9 Conclusions

Un cop acabat el treball, estic satisfeta al poder dir que he pogut crear un model que prediu si la persona amb cirrosis hepàtic morirà, sobreuirà o sobreuirà amb trasplantament.

Després de desenvolupar el model, he pogut arribar a la conclusió que els millors models creats són un SVM o un EBM, segons la prioritat del client.

No obstant això, és cert que els models no tenen tanta exactitud com esperava, ja que hi ha una petita mostra de sobreajustament a les dades d'entrenament. A més a més, es pot notar com les prediccions no funcionen correctament a l'hora de predir una classe concreta. Això és deu a les poques dades que hi ha al conjunt de dades inicial i pot ser un problema o bé de l'oversampling al crear les noves instàncies o bé del test, ja que la proporció de persones en aquesta classe és minoritària.

Això es podria solucionar agafant un conjunt de dades més gran. És cert que desenvolupant aquest treball m'he trobat en diversos problemes respecte a les poques dades i la baixa qualitat d'aquestes.

Finalment, aplicant el hierarchical clustering i el KMeans hem pogut observar com no es poden crear clústers que corresponguin a les diferents classes de la categoria objectiu.

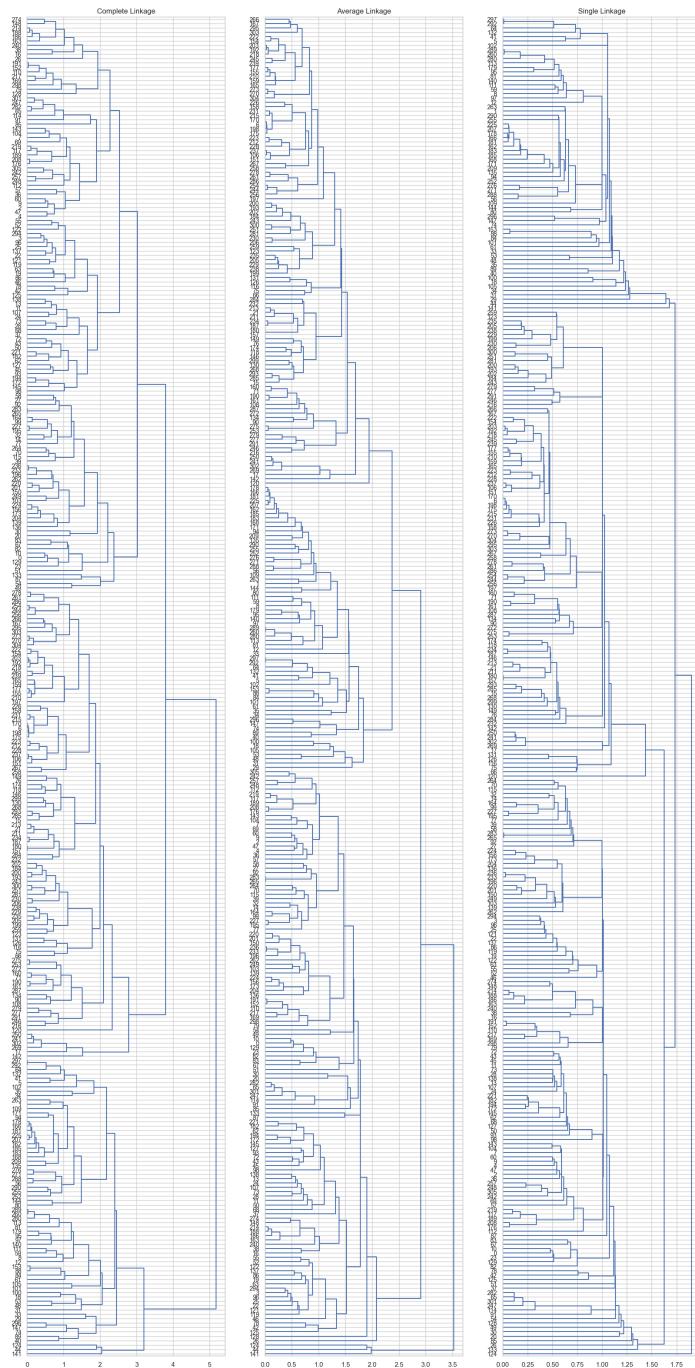


Figura 168: Hierarchical Clustering

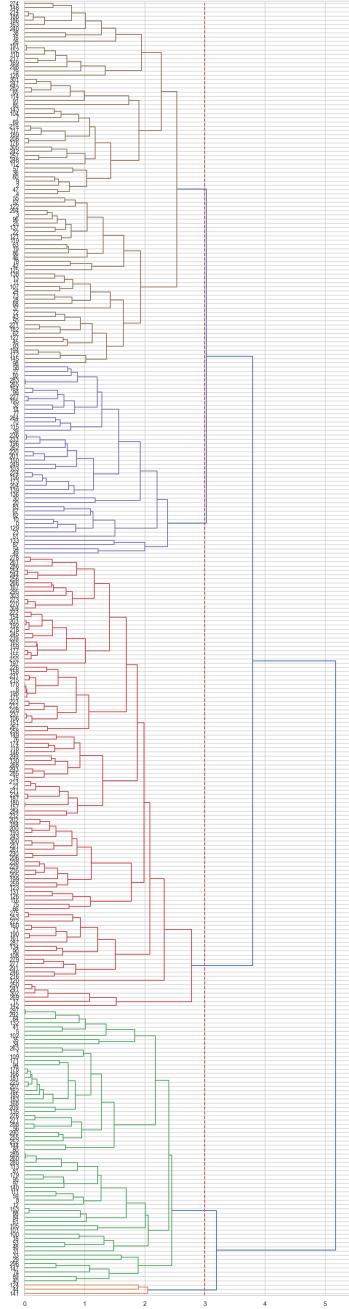


Figura 169: Hierarchical Clustering