UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Laboratori 1

Inteligencia artificial

Introducció a l'Aprenentatge Automàtic

Autor

Artur Aubach Altes

Grup 12

Profesor: Jordi Luque Serrano

Quadrimestre Tardor 2023/2024

TAULA DE CONTINGUTS

1. IDENTIFICACIÓ DEL PROBLEMA 4
1.1 Anàlisis dels models 4
1.1.1 KNN 4
1.1.2 Arbre de decisions 4
1.1.3 SVM 5
1.2 Primera visio 5
2. ANÀLISIS I PREPROCESSAT DE DADES 7
2.1 Preparació de Dades Abans de la Partició 7
2.1.1 Anàlisi univariant 7
2.1.1.1 Categoriques i binarias 7
2.1.1.2 Numeriques 13
2.1.2 Estudi de balanceig de classes 26
2.1.3 Detecció de missings 27
2.1.4 Detecció e imputació d'outliers 27
2.1.5 Feature engineering 31
2.1.5.1 Construction 31
2.1.5.2 Selection 31
2.2 Gestió de Dades Post-Partició 32
2.2.1 Partició 32
2.2.2 Tractament de missings 32
2.2.2.1 Determinació del Millor Mètode d'Imputació 32
2.3.2.2 Resultats de l'Avaluació 33
2.2.3 Tractament de balanceig 34
2.2.3.1 Anàlisi dels Mètodes de Balanceig 34
2.2.3.2 Conclusió 35
2.2.4. Codificació de variables 35
2.3.5 Análisis post processament 37
3. PREPARACIÓ DE VARIABLES 45
3.1 Normalització i escalat 45
3.2 Correlacions entre variables numèriques 50
3.3 Anàlisi de variables categòriques i variable objectiu 51
3.3.1 Interpretació dels Resultats 51
3.4 Eliminació de variables redundants o sorolloses 52
3.5 Estudi de dimensionalitat amb PCA. 52
3.5.1 Explicació dels resultats 53
3.5.2 Justificació de no aplicar la reducció amb PCA 54
3.6 Ultimes consideracions 55
3.7 Resum de les dedas 55
4. Definició de models 56
4.1 Definició de mètriques 56
4.2 Primer mode (KNN) 57

4.2.1 Motivació 57
4.2.2 Discussió dels hiperparàmetres disponibles, i dels valors usats 57
4.2.3 Entrenament 58
4.2.4 Anàlisi de resultats 59
4.3 Segon model (Arbre de decisions) 60
4.3.1 Motivació 60
4.3.2 Discussió dels hiperparàmetres disponibles, i dels valors usats 60
4.3.3 Entrenament 61
4.3.4 Anàlisi de resultats 63
4.2 Tercer model (SVM) 64
4.4.1 Motivació 64
4.4.2 Discussió dels hiperparàmetres disponibles, i dels valors usats 65
4.4.3 Entrenament 65
4.4.4 Anàlisi de resultats 66
4.3 Comparació 66
5. SELECCIÓ DE MODEL 68
5.1 Descripció del model triat 68
5.2 Anàlisi de les limitacions i capacitats del model. 69
6. MODEL CARD 70
7. BONUS 2 (clusters) 72
7.1 Predir la supervivència dels pacients, tractant la variable 'Status' 72
7.2 Clusters genèrics 77

1. IDENTIFICACIÓ DEL PROBLEMA

El propòsit d'aquest treball és desenvolupar un model predictiu per a pacients amb cirrosi hepàtica, utilitzant la base de dades “Cirrhosis Patient Survival Prediction” (https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1), que inclou 17 característiques clíniques. L'objectiu principal és predir la supervivència dels pacients, tractant la variable 'Status' com a variable objectiu.

1.1 Anàlisis dels models

Per la predicció, es consideraran tres tipus de models de classificació: KNN (K-Nearest Neighbors), Arbre de Decisions, i SVM (Support Vector Machine). Aquesta fase inicial es centrarà en una anàlisi teòrica detallada de cada model. Aquesta comprensió inicial serà crucial per determinar quins models són més adequats per a aquesta aplicació específica. Tot i que aquesta serà una aproximació preliminar, l'elecció dels models i la configuració dels seus paràmetres es refiniran continuament al llarg de la fase experimental del projecte.

1.1.1 KNN

KNN és un algoritme de classificació supervisat que opera sota el principi que punts de dades similars tendeixen a estar propers. En el context de la predicció de la supervivència de pacients amb cirrosi hepàtica, KNN pot ser eficaç perquè permet una aproximació intuïtiva a patrons complexos en dades clíniques, considerant la similitud entre pacients. Tot i que el KNN gestiona bé les dades numèriques, les característiques categòriques poden requerir codificació per millorar la seva aplicabilitat. La normalització és fonamental en KNN, donada la seva dependència de la proximitat entre els punts de dades.

1.1.2 Arbre de decisions

Els arbres de decisions són models de classificació supervisat, intuïtius que es basen en la segregació successiva de dades en grups més petits, fent decisions basades en característiques específiques. Aquesta aproximació és útil en medicina, ja que pot reflectir el procés de presa de decisions clíniques. Un avantatge clau és la seva

capacitat per manejar dades categòriques i numèriques directament, fent-lo ideal per a conjunts de dades clíniques variats. A més, no requereixen normalització ni escalat de dades, ja que la seva estructura de decisió no es basa en la magnitud dels valors de les característiques.

1.1.3 SVM

SVM és un potent model de classificació supervisat que busca el millor hiperplà que separa les diferents classes en un espai de característiques d'alta dimensió. Aquesta separació es realitza maximitzant el marge entre les classes més properes. En el context de la predicció de supervivència, SVM és útil per la seva capacitat de gestionar eficaçment espais de característiques complexes i de grans dimensions. Encara que SVM pot treballar amb dades numèriques i categòriques, la codificació de les últimes pot ser necessària. La normalització i l'escalat de dades són importants per assegurar que totes les característiques contribueixen equitativament a la formació del model.

1.2 Primera visio

Després d'analitzar les característiques específiques de cada model en la secció "1.1 Anàlisi dels models", s'ha observat que tots ells són compatibles amb la tècnica de validació creuada. Aquesta compatibilitat és particularment rellevant, ja que la validació creuada és una estratègia clau en la modelització predictiva. Aquesta metodologia permet una avaluació rigorosa del rendiment del model, dividint el conjunt de dades en diverses parts per a l'entrenament i la validació. Aquesta aproximació ajuda a garantir que el model és capaç de generalitzar bé a noves dades, una qualitat essencial en l'àmbit mèdic on les prediccions han de ser tant precises com fiables.

Amb aquesta consideració en ment, la següent taula resumeix les característiques principals dels models:

Model	Accepta Cross-Validation	Accepta Dades Categòriques Sense Codificació	Necessitat de Normalització	Necessitat d'Escalar Dades
KNN	SÍ	NO *(requereix codificació)*	NO	SÍ (recomanat)
Arbre de Decisions	SÍ	SÍ	NO	NO
SVM	SÍ	NO *(requereix codificació)*	SÍ	SÍ

NOTA IMPORTANT: *Encara que teòricament els Arbres de Decisió accepten variables categòriques, en la pràctica, la llibreria sklearn no disposa de cap model d'arbre de decisions que les accepti. No em vaig adonar compte d'aquest detall fins a la fase final del entrenament del model moment. Per això, durant el desenvolupament, es pot observar que vaig intentar mantenir dues bases de dades: una amb variables numèriques i categòriques, i una altra només amb numèriques; una distinció que, en realitat, no era necessària.*

2. ANÀLISIS I PREPROCESSAT DE DADES

L'objectiu principal d'aquesta secció és dur a terme un anàlisi detallat i el preprocessament de les dades per garantir la millor qualitat possible per a la construcció dels models. Aquest procés inclou la neteja de dades per tractar valors perduts o erronis, l'anàlisi exploratòria per entendre la distribució i les relacions entre les característiques, la codificació de dades categòriques per a la seva utilització en models com KNN i SVM, la normalització i l'escalat per assegurar la comparabilitat i eficàcia dels models, i, si és necessari, la reducció de la dimensionalitat per optimitzar l'eficiència del model.

2.1 Preparació de Dades Abans de la Partició

Aquesta secció es centra en els processos essencials de preparació de dades abans de realitzar la partició del conjunt de dades per a la modelització.

2.1.1 Anàlisi univariant

L'anàlisi univariant en aquesta fase del preprocessament de dades s'ha realitzat de manera separada per a dues categories de dades: les dades categòriques i binàries, i les dades numèriques.

2.1.1.1 Categoriques i binarias

En aquesta etapa de l'anàlisi univariant, es va iniciar examinant quins eren els atributs categòrics i binaris del conjunt de dades.

```
Python
Number of categorical and binaria variables:  11
**************************************************************************
Categorical variables column name:  ['Status', 'Drug', 'Sex', 'Ascites',
'Hepatomegaly', 'Spiders', 'Edema', 'Cholesterol', 'Copper', 'Tryglicerides',
'Platelets']
```

Inicialment, es van identificar 11 variables com a categòriques o binàries.

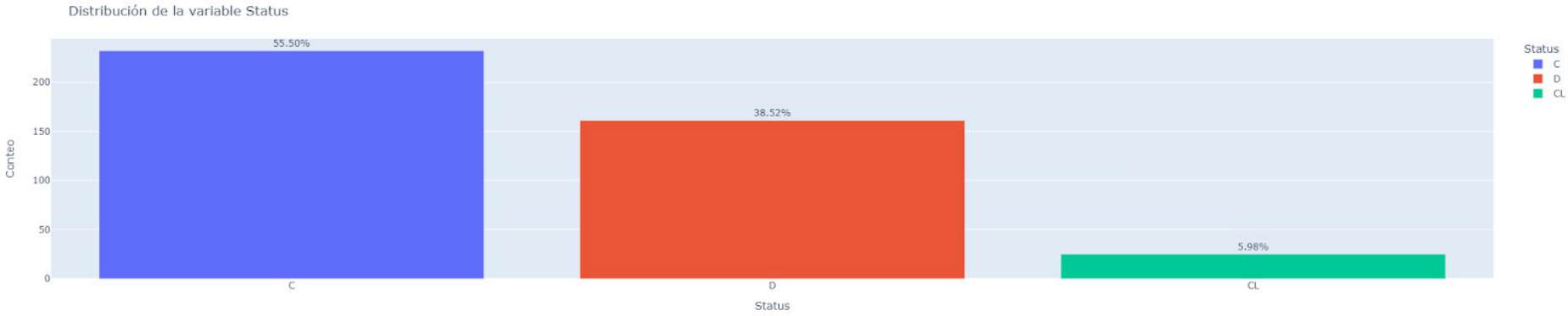
No obstant això, es va observar una inconsistència, ja que alguns atributs classificats com a categòrics ('Cholesterol', 'Copper', 'Tryglicerides', 'Platelets') eren, de fet, numèrics, però estaven representats com a cadenes de text (strings).

```python
Cholesterol es: <class 'str'>
Copper es: <class 'str'>
Tryglicerides es: <class 'str'>
Platelets es: <class 'str'>
```

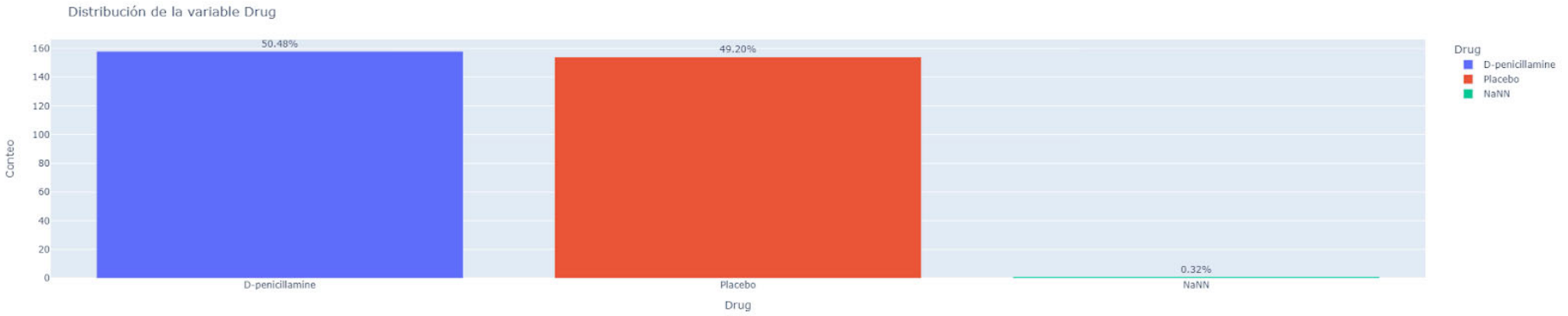
Aquesta observació va portar a la decisió de convertir aquests atributs a format numèric.

```python
Number of categorical and binaria variables:  7
********************************************************************************
Categorical variables column name:  ['Status', 'Drug', 'Sex', 'Ascites',
'Hepatomegaly', 'Spiders', 'Edema']
```

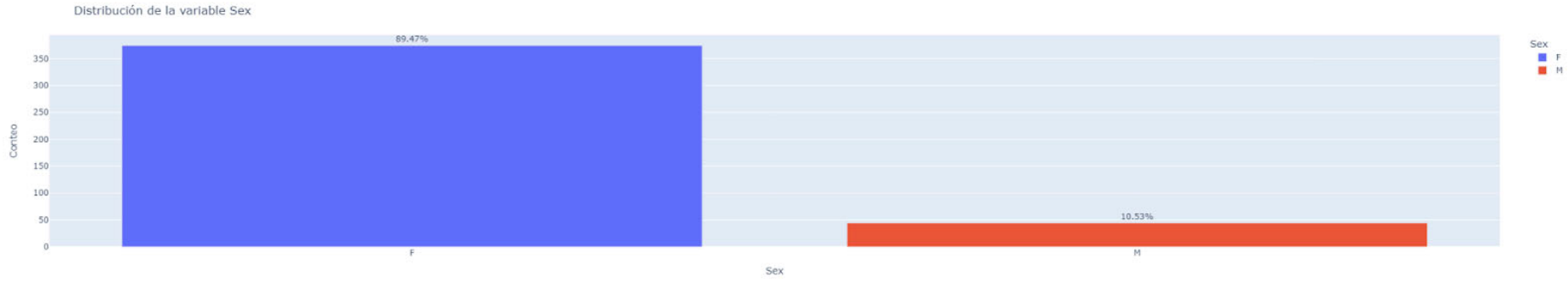
Després d'aquesta correcció, el nombre de variables categòriques i binàries es va reduir a 7: ['Status', 'Drug', 'Sex', 'Ascites', 'Hepatomegaly', 'Spiders', 'Edema'].



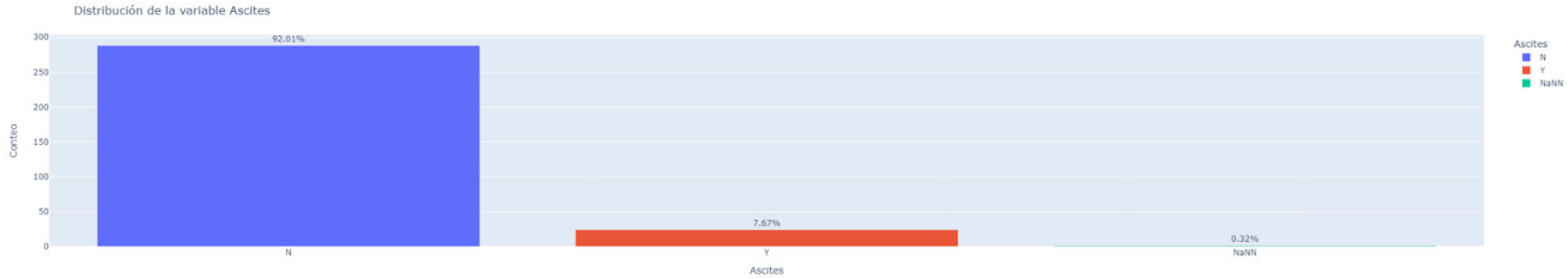
Status: Aquesta variable indica l'estatus del pacient. Pot ser 'C' (censurat) -**55,50%**, 'CL' (censurat degut a transplantament de fetge) **-5,98%**, o 'D' (mort) **-38,52%**. Aquesta classificació és crucial per entendre l'evolució de cada pacient en l'estudi.



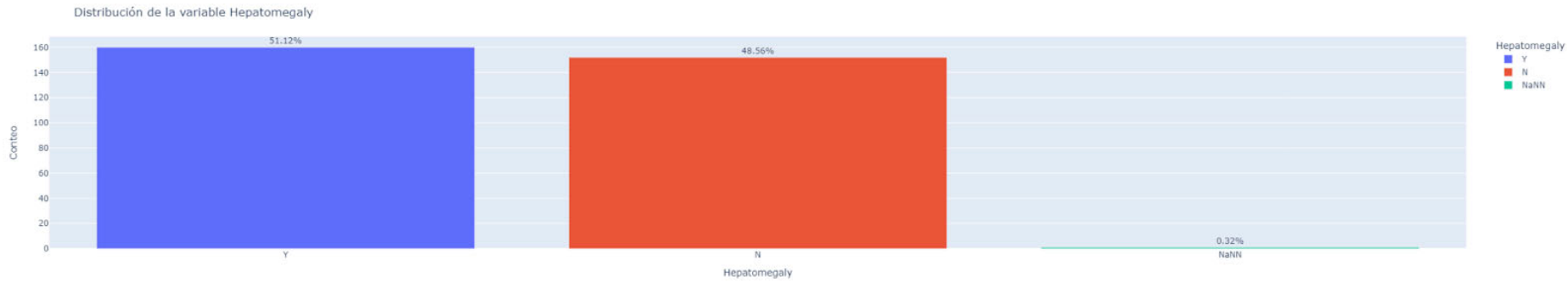
Drug: Tipus de medicament administrat al pacient, que pot ser D-penicil·lamin **-50,48%** o placebo **-59,20%**. Aquesta variable és important per determinar l'efecte del tractament en la supervivència dels pacients. I de missings **-0,32%**



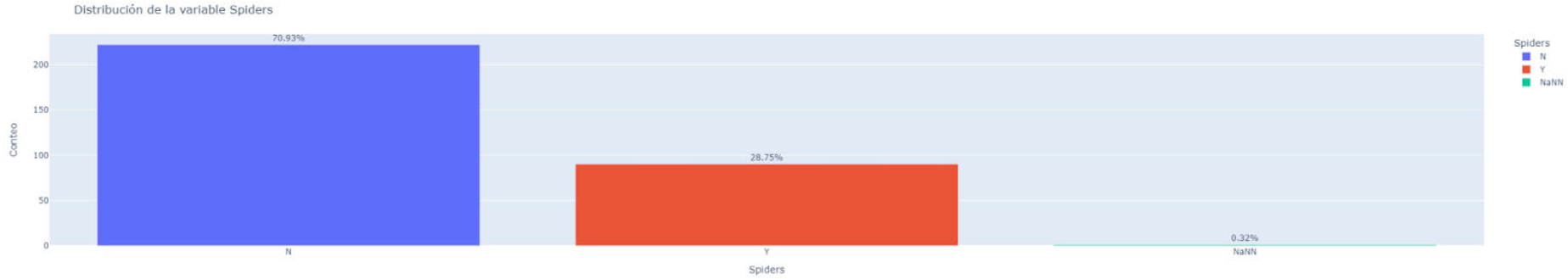
Sex: Gènere del pacient, que pot ser M (masculí) **-92,02%** o F (femení) **-7,67%**. El gènere pot ser un factor rellevant en l'anàlisi de la progressió de la malaltia.



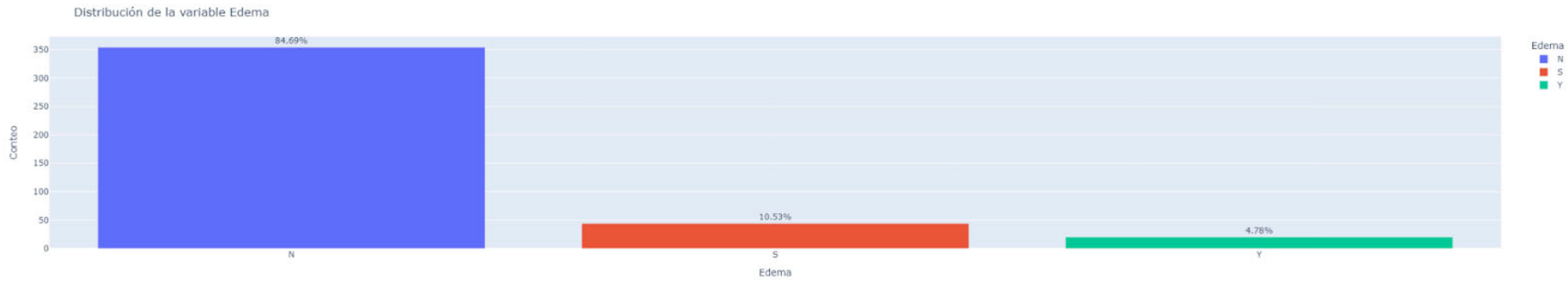
Ascites: Presència d'ascites, indicada com 'N' (No) **-51,12%** o 'Y' (Sí) -. L'ascites és un símptoma clínic significatiu en pacients amb cirrosi hepàtica. I de missings **-0,32%.**



Hepatomegaly: Presència d'hepatomegàlia, indicada com 'N' (No) o 'Y' (Sí). L'hepatomegàlia pot ser un indicador de l'estat de la malaltia hepàtica. I de missings **-0,32%.**



Spiders: Presència d'aranyes vasculars, indicada com 'N' (No) o 'Y' (Sí). Aquesta característica pot estar associada amb malalties hepàtiques avançades. I de missings **-0,32%.**



Edema: Aquesta variable descriu la presència d'edema. Pot ser 'N' (sense edema i sense tractament diürètic per a l'edema), 'S' (edema present sense diürètics, o edema resolt per diürètics) o 'Y' (edema malgrat la teràpia diürètica). L'edema és un símptoma clau en el maneig clínic de la cirrosi. I de missings **-4,78%.**

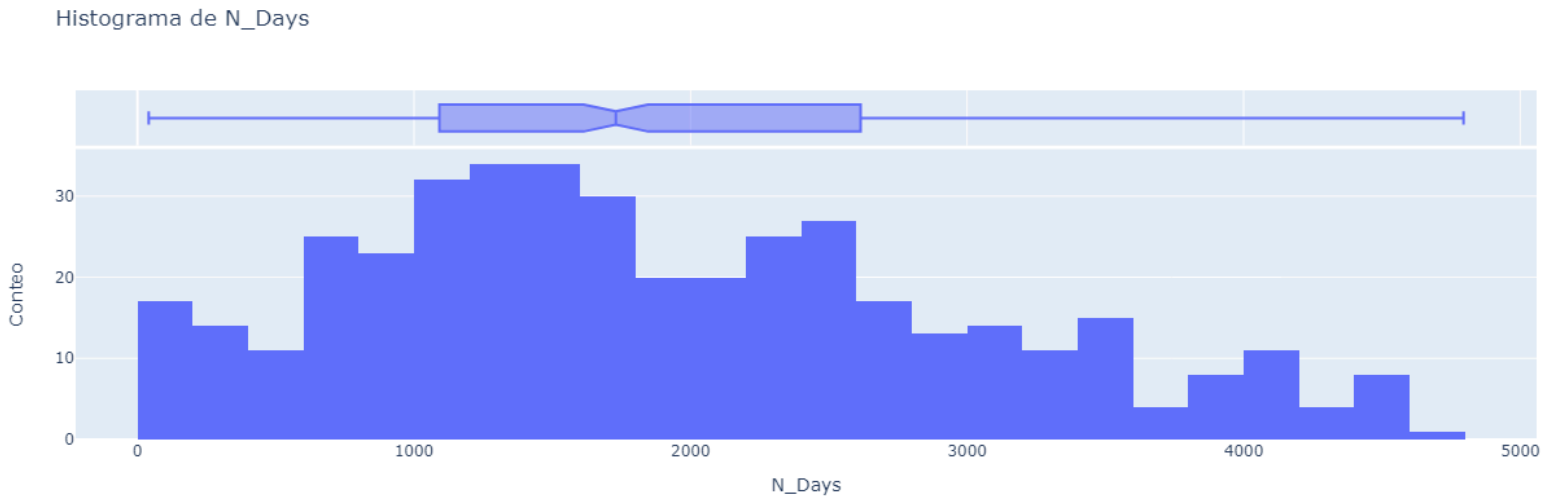
Aquesta informació s'ha tret de UC Irvine Machine Learning Repository que es d'on s'ha extret la base de dades.

2.1.1.2 Numeriques

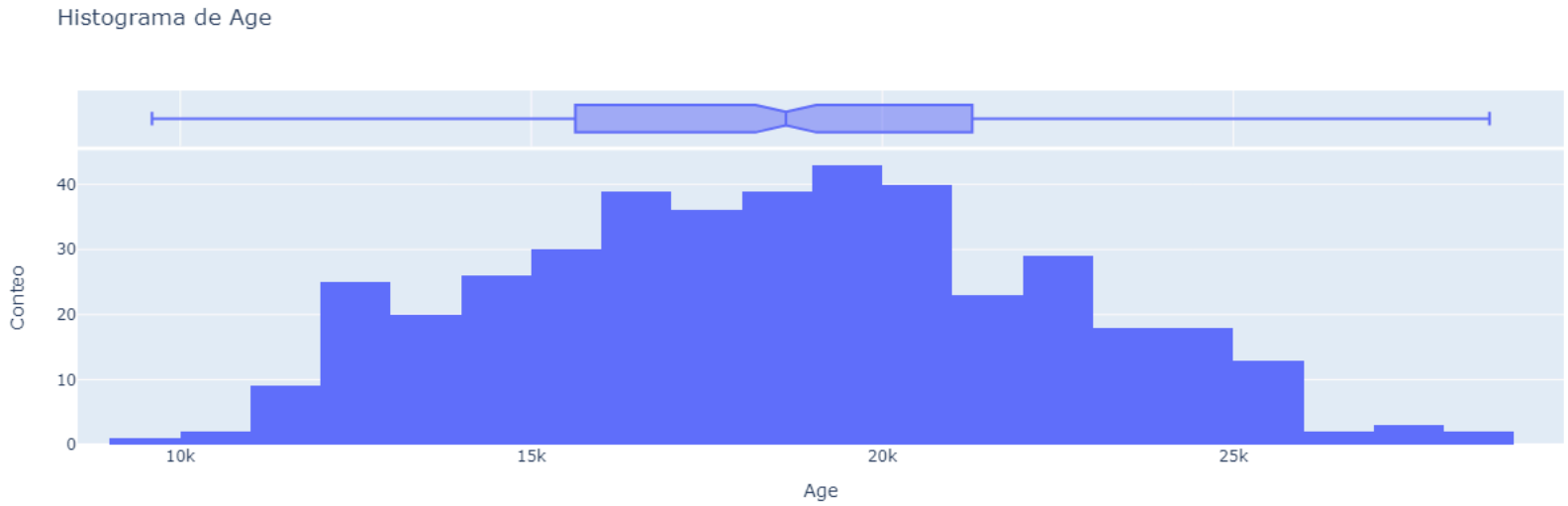
```
Python
Number of numerical variables:  13
********************************************************************************
Numerical  Variables  Column:   ['ID',  'N_Days',  'Age',  'Bilirubin',  'Cholesterol',
'Albumin',  'Copper',  'Alk_Phos',  'SGOT',  'Tryglicerides',  'Platelets',  'Prothrombin',
'Stage']
```

En l'anàlisi univariant de les dades numèriques, s'han identificat 13 variables: ['ID', 'N_Days', 'Age', 'Bilirubin', 'Cholesterol', 'Albumin', 'Copper', 'Alk_Phos', 'SGOT', 'Tryglicerides', 'Platelets', 'Prothrombin', 'Stage'].

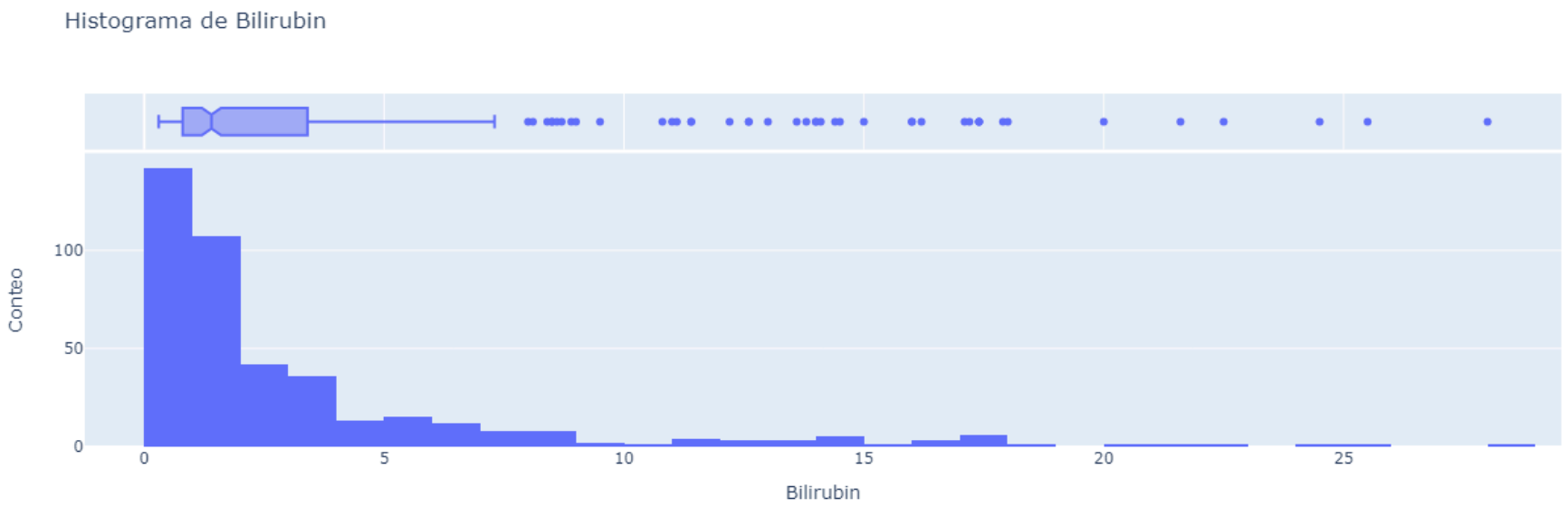
S'ha observat que l'atribut "ID", sent un identificador únic, no aporta cap valor analític, per tant, s'ha decidit excloure'l de l'anàlisi.



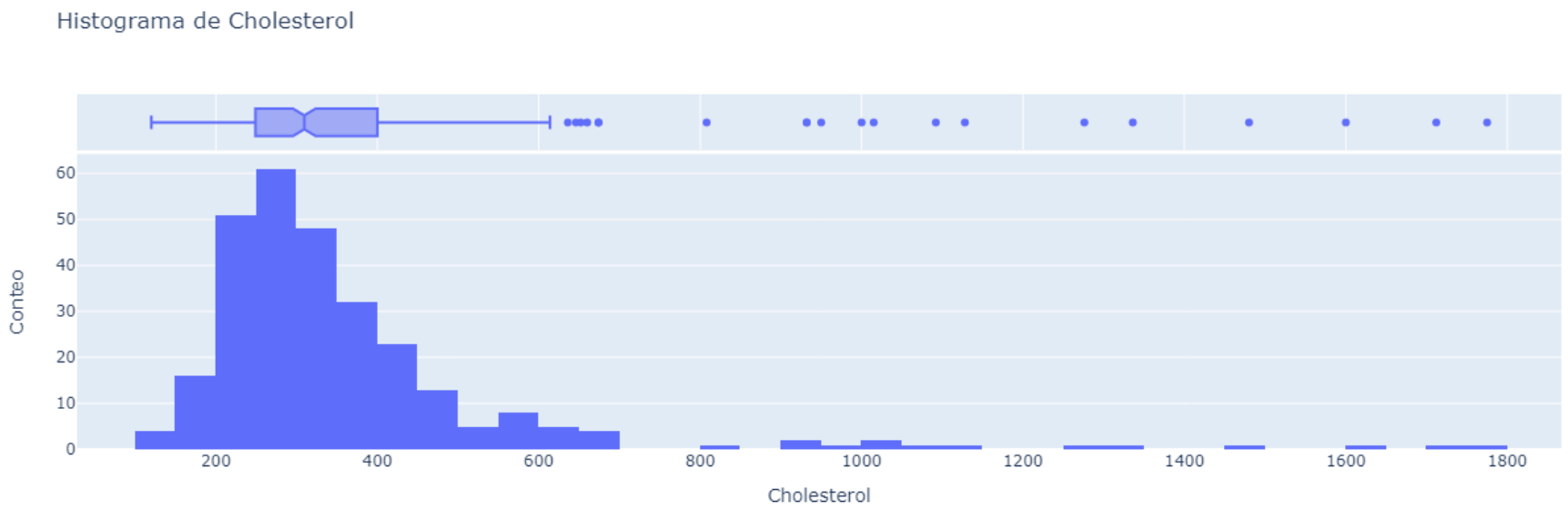
N_Days: Nombre de dies des de la inscripció fins a la mort, el trasplantament o el temps d'anàlisi de l'estudi. L'histograma mostra una distribució amb una concentració de valors cap a la part inferior de l'escala de dies, suggerint que la majoria dels pacients són registrats o tenen esdeveniments (mort, trasplantament) en un període més curt després del registre. La presència de dades a la cua dreta de la gràfica indica que hi ha pacients que han sobreviscut o han estat seguits per un període de temps significativament llarg. La caixa de bigotes sobre l'histograma indica una mediana relativament baixa comparada amb la mitjana, suggerint una distribució asimètrica amb una cua cap a la dreta.



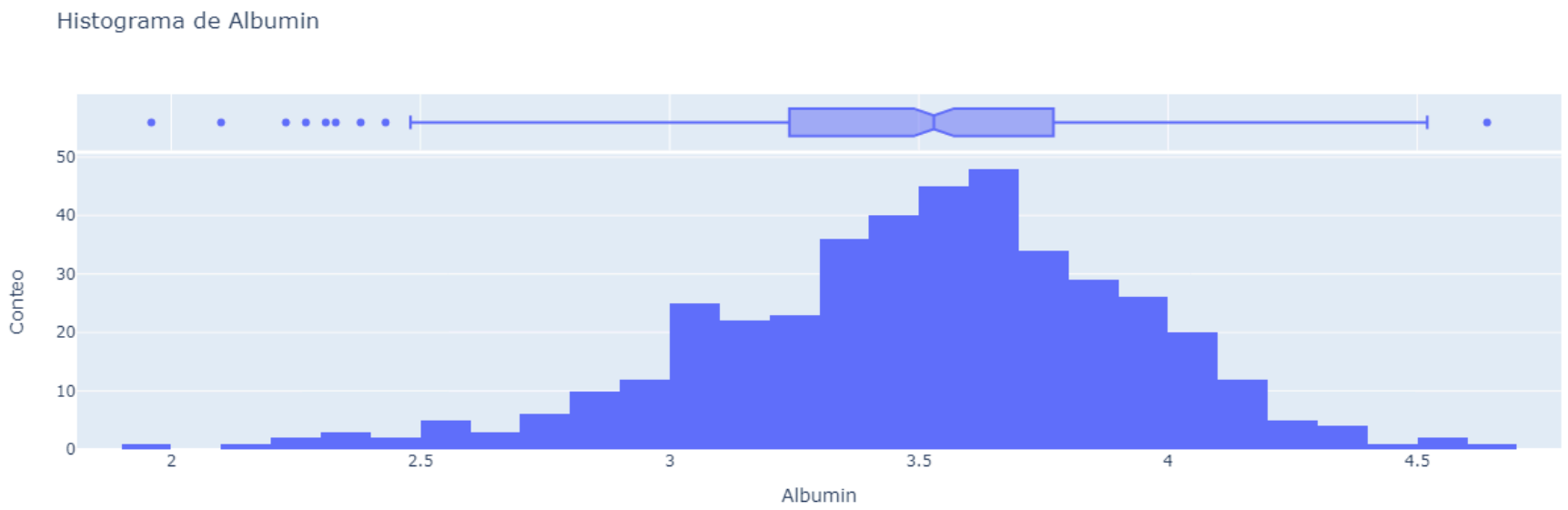
Age: Edat dels pacients en dies. Els dies d'edat dels pacients mostren una distribució relativament simètrica amb una lleugera cua a la dreta. La majoria de pacients es concentren en un rang intermedi d'edat, amb uns quants casos que mostren edats més altes que la mitjana.



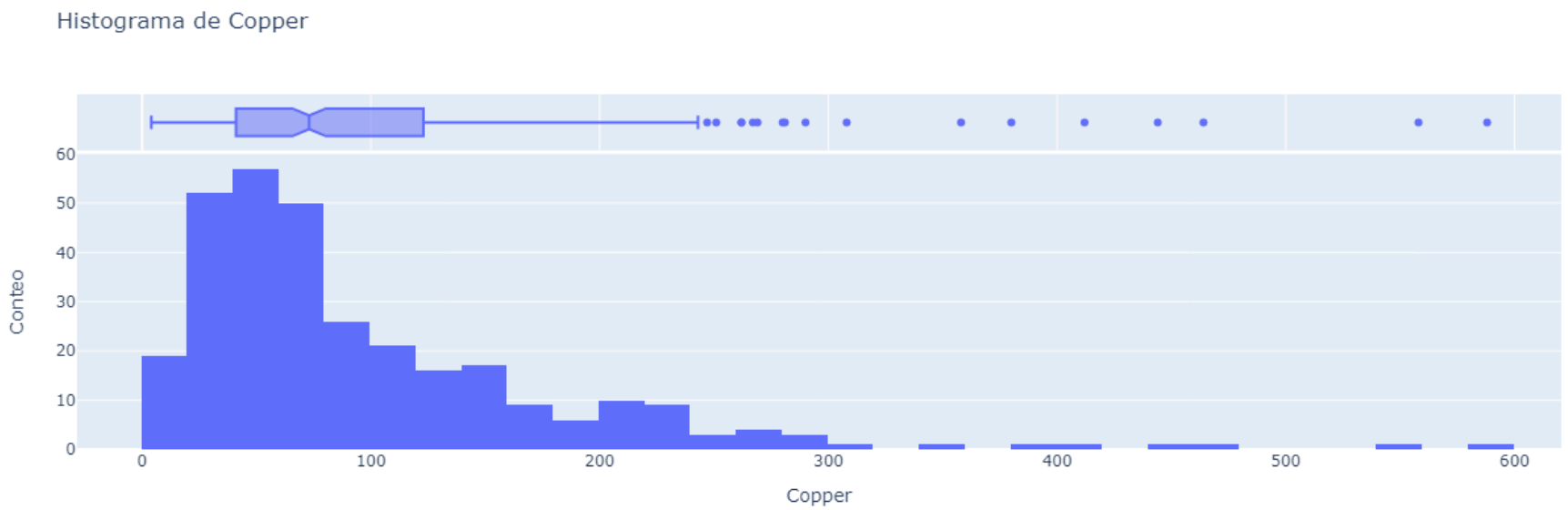
Bilirubin: Nivell de bilirubina en sang. L'histograma de bilirubina mostra una distribució molt sesgada cap a la dreta, amb la majoria de valors concentrats a l'extrem inferior de l'escala i uns pocs valors extremadament alts que s'estenen com a punts fora dels bigotes de la caixa, indicant posibles outliers.



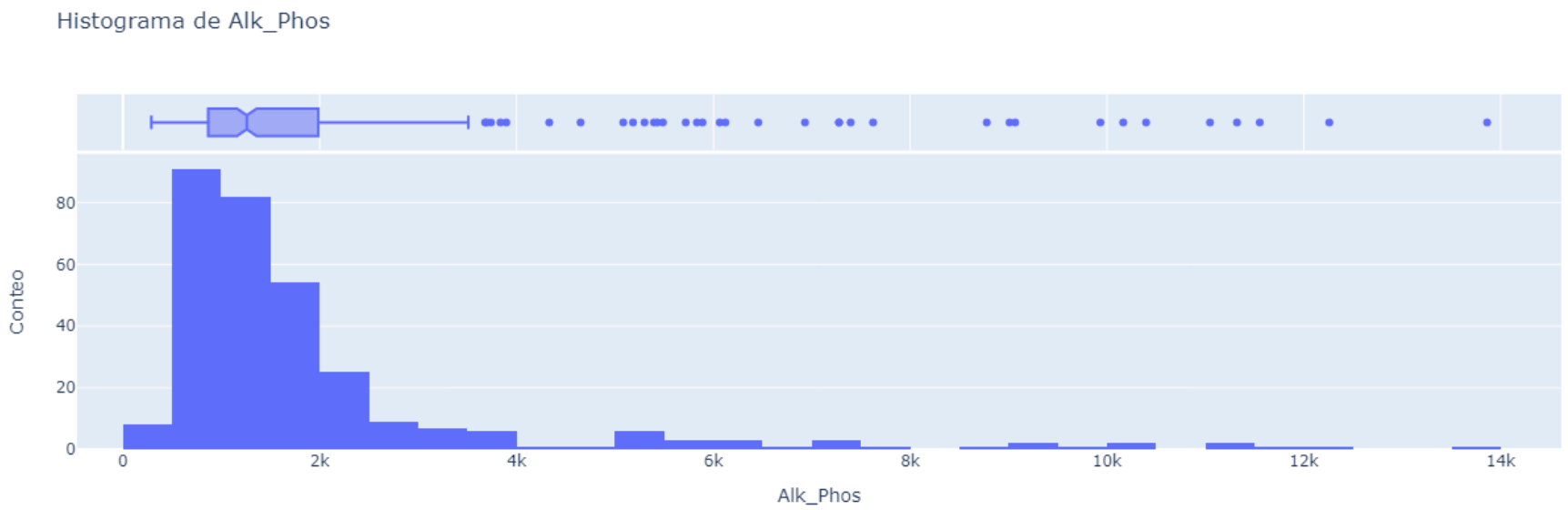
Cholesterol: Nivell de colesterol en sang. L'histograma de colesterol presenta una distribució sesgada cap a la dreta, amb la majoria de pacients tenint nivells més baixos i uns quants amb valors extremadament alts.



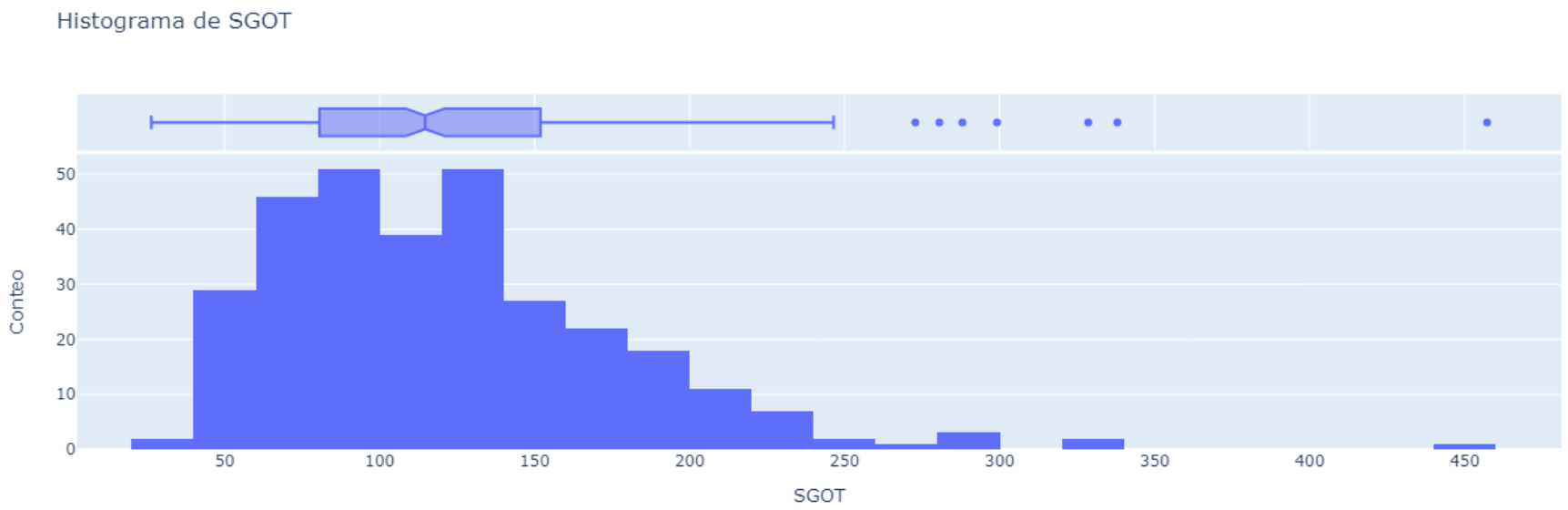
Albumin: Nivell d'albumina en sang. La distribució dels nivells d'albumina mostra una forma més simètrica, amb la majoria dels valors agrupats al voltant del centre de l'escala. No obstant això, hi ha una lleugera asimetria cap a la dreta. El diagrama de caixa indica que la mediana està centrada dins del rang interquartílic, i hi ha uns pocs posibles outliers a la part superior de la distribució.



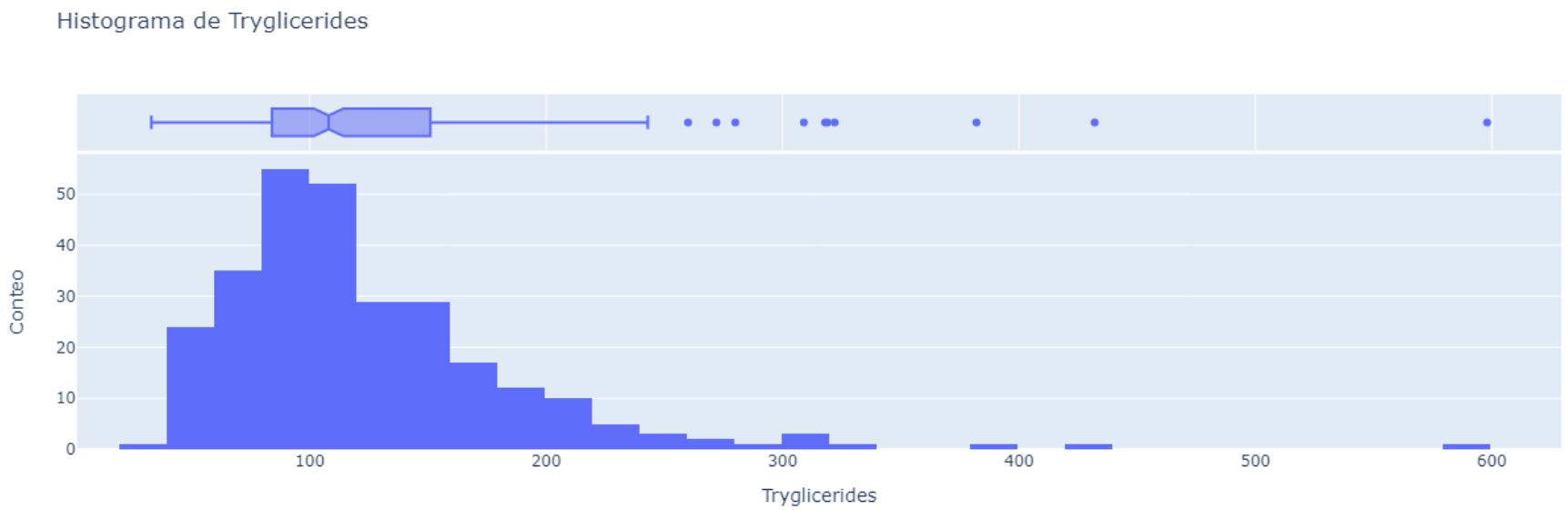
Copper: Nivell de coure en orina. Aquest histograma també mostra una distribució sesgada cap a la dreta, amb una concentració de valors baixos i una presència notable de posibles outliers a la dreta, indicant que alguns pacients presenten nivells molt alts de coure en orina, el que podria ser rellevant per a l'avaluació de la malaltia hepàtica.



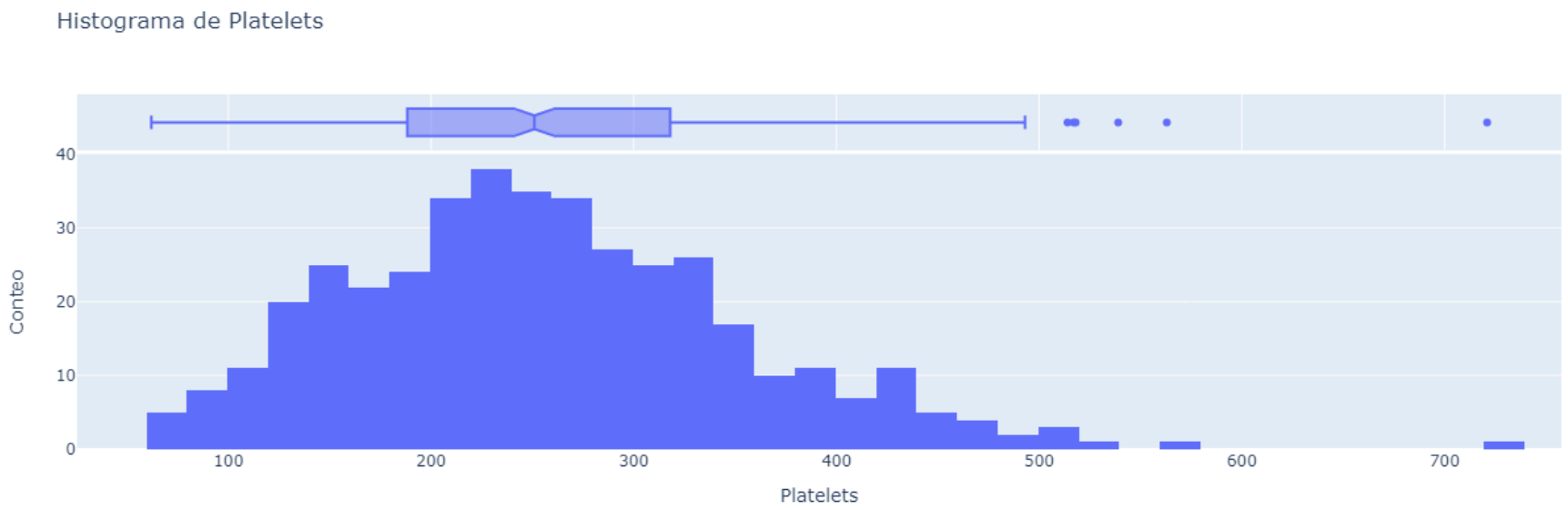
Alk_Phos: Nivell de fosfatasa alcalina en sang. La distribució d'Alkaline Phosphatase (Alk_Phos) és altament asimètrica, amb una concentració de valors baixos i un nombre significatiu de posibles outliers, com es mostra en el diagrama de caixa adjunt a l'histograma. Això pot reflectir una varietat en la gravetat de la malaltia hepàtica entre els pacients.



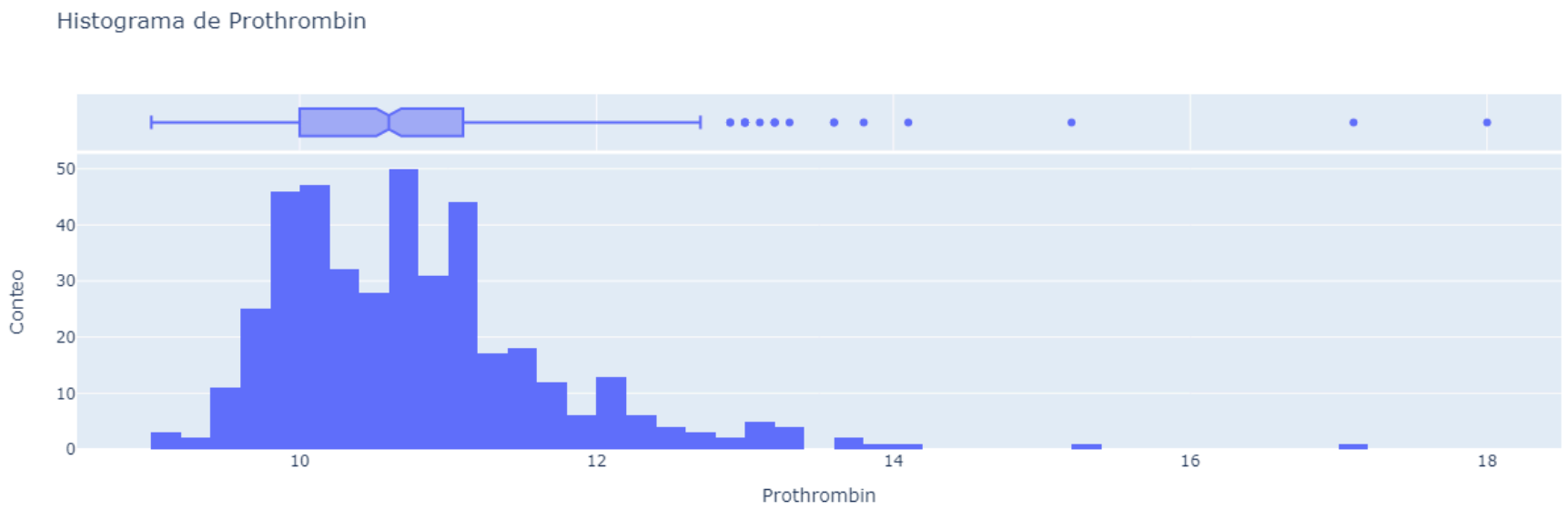
SGOT: Nivell d'SGOT en sang. Aquest histograma mostra una distribució amb una concentració de dades en els valors més baixos i una cua llarga cap a la dreta. El diagrama de caixa indica presència de posibles d'outliers.



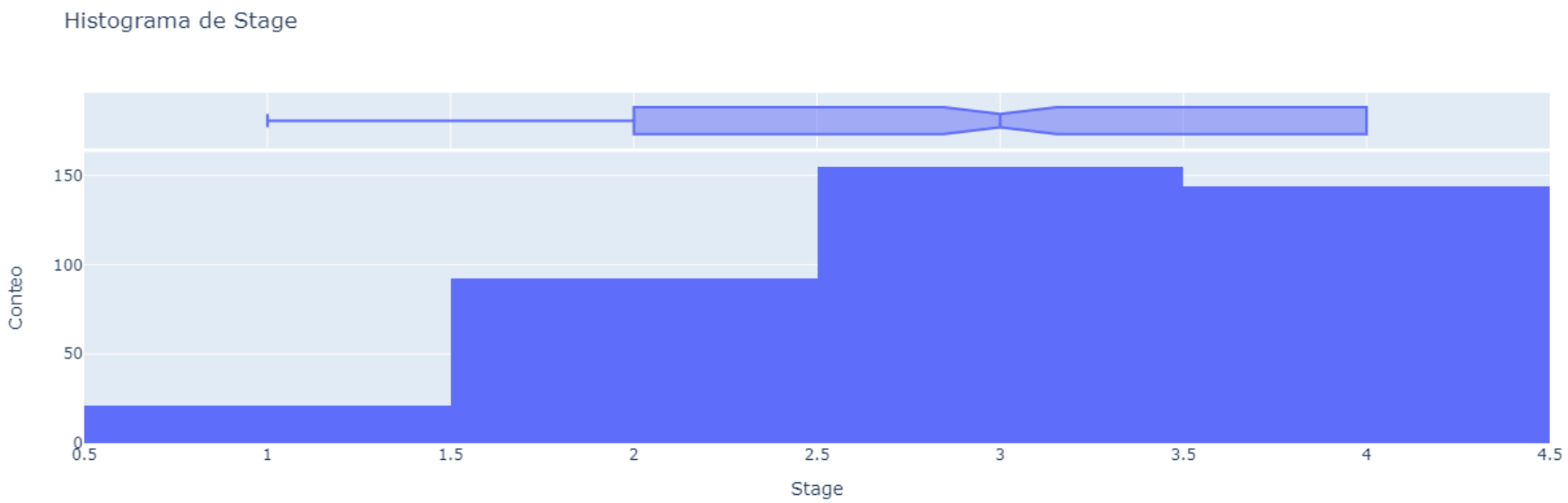
Trygliceridos: Nivell de triglicèrids en sang. Aquest histograma mostra una distribució sesgada cap a la dreta, amb la majoria de dades concentrades en valors més baixos i alguns posibles outliers. Això indica que la majoria dels pacients tenen nivells de triglicèrids dins d'un rang normal amb uns pocs amb nivells elevats.



Platelets: Nombre de plaquetes per cúbic ml/1000. L'histograma mostra una distribució amb una forma que s'assembla a la normalitat, però amb una lleugera cua cap a la dreta, suggerint la presència d'alguns pacients amb comptes de plaquetes molt alts.



Prothrombin: Temps de protrombina en segons. La distribució del temps de protrombina mostra un patró amb múltiples modes, possiblement indicant diferents grups o condicions dins de la població estudiada.



Stage: Etapa histològica de la malaltia. Aquest histograma revela que una gran proporció de pacients estan en una etapa avançada de la malaltia (etapa 4), suggerint que aquest conjunt de dades pot incloure molts pacients amb cirrosi hepàtica greu.

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Trygliceride	Platelets	Prothrombin	Stage
count	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
mean	1714.09	14570.02	57.44	467.43	54.88	165.31	2837.12	165.26	180.76	287.86	60.81	53.74
std	1491.85	9389.60	145.97	535.34	146.73	193.56	4515.53	146.19	184.46	207.65	143.59	144.76
min	41.00	418.00	0.30	120.00	0.42	4.00	289.00	26.35	33.00	62.00	1.02	0.88
25%	924.06	8152.46	1.25	245.11	2.92	65.06	731.62	74.62	79.47	165.96	9.75	1.75
50%	1417.34	17088.93	3.31	296.75	3.51	91.63	1619.50	118.63	116.35	254.01	10.67	3.01
75%	2091.71	19289.12	10.31	377.13	3.99	169.75	2022.09	191.92	183.75	340.25	12.82	4.00
max	4795.00	28650.00	418.00	1775.00	418.00	588.00	13862.40	457.25	598.00	721.00	416.00	412.00

La taula presenta estadístiques descriptives de les variables numèriques. La variable N_Days mostra una gran variabilitat, amb una desviació estàndard (std) prou elevada en comparació amb la mitjana (mean), indicant una ampla dispersió en el nombre de dies. Similarment, la variable Age també presenta una alta desviació estàndard relativa a la seva mitjana, assenyalant una variabilitat significativa en les edats dels pacients. Els valors elevats de std en variables com Bilirubin, Cholesterol, i Copper suggereixen heterogeneïtat dins del conjunt de dades.

Observant aquestes estadístiques, es fa evident que les dades requereixen normalització o estandardització, especialment per a l'ús en models com KNN o SVM, que són sensibles a l'escala de les dades, com es va esmentar a la secció "1.2 Primera Visió".

2.1.2 Estudi de balanceig de classes

```
Python
Status
C     232
D     161
CL     25
Name: count, dtype: int64
```

En la secció "2.1.2 Estudi de balanceig de classes", s'ha detectat un desequilibri en la variable objectiu "Status". Amb 232 casos censurats (C), 161 morts (D) i només 25 censurats per trasplantament de fetge (CL), es manifesta una clara disparitat. Aquest desequilibri pot afectar la capacitat predictiva dels models, ja que poden

tendir a favorir la classe majoritària. Per tant, serà essencial considerar l'aplicació de mètodes de balanceig de classes.

2.1.3 Detecció de missings

En aquesta secció, es va abordar l'anàlisi dels valors absents en el nostre conjunt de dades, enfocant-nos específicament en els atributs numèrics. Cal recordar que, per a les variables categòriques i booleanes, ja es va realitzar una anàlisi visual en l'apartat "2.1.1.1. A continuació" on es van poder observar els missings. A continuació, es va centrar l'atenció en l'examen dels missings en les variables numèriques:

```
Python
La variable ID no té valors absents.
La variable N_Days no té valors absents.
La variable Age no té valors absents.
La variable Bilirubin no té valors absents.
La variable Cholesterol té 134 valors absents, el que representa un 32.06%.
La variable Albumin no té valors absents.
La variable Copper té 108 valors absents, el que representa un 25.84%.
La variable Alk_Phos té 106 valors absents, el que representa un 25.36%.
La variable SGOT té 106 valors absents, el que representa un 25.36%.
La variable Tryglicerides té 136 valors absents, el que representa un 32.54%.
La variable Platelets té 11 valors absents, el que representa un 2.63%.
La variable Prothrombin té 2 valors absents, el que representa un 0.48%.
La variable Stage té 6 valors absents, el que representa un 1.44%.
```

Davant d'aquesta situació, es va considerar que la millor opció era l'aplicació de mètodes d'imputació de valors absents que fossin compatibles amb tipus de dades mixtes (booleanes, numèriques i categòriques). Aquests mètodes van ser seleccionats per ser capaços de tractar de manera eficient la diversitat dels tipus de dades presents en el conjunt, per assegurar una imputació acurada i millorar la integritat de les dades per a anàlisis posteriors.

2.1.4 Detecció e imputació d'outliers

Bilirubina (0.30-418.00 mg/dl):

- El rang normal de bilirubina total per adults és generalment entre 0.2 i 1.3 mg/dL, segons Cleveland Clinic.
- Tot i que un valor de 418 mg/dl és extremadament alt, hi ha condicions mèdiques que podrien causar una elevació significativa de la bilirubina. Per exemple, problemes greus de fetge, obstruccions dels conductes biliars o una destrucció accelerada de glòbuls vermells podrien resultar en nivells elevats de bilirubina. No obstant això, és important destacar que aquest valor és molt poc comú i seria indicatiu de condicions mèdiques severes.

https://my.clevelandclinic.org/health/diagnostics/17845-bilirubin

Colesterol (120.00-1775.00 mg/dl):

- Normalment, els nivells totals de colesterol saludables per a adults són de 125 a 200 mg/dL. Un nivell total de colesterol per sobre de 240 mg/dL ja es considera alt. Així, un valor de 1775 mg/dL és molt fora del rang normal i és extremadament inusual.
- Aquests nivells podrien ser possibles en casos d'hipercolesterolemia severa o altres trastorns metabòlics, però són excepcionals i indicarien una condició mèdica seriosa que requeriria atenció mèdica immediata.

https://www.mayoclinic.org/tests-procedures/cholesterol-test/about/pac-20384601

Albumina (0.42-418.00 g/dL):

- El rang normal d'albumina en la sang d'un adult és de 3.4 a 5.4 g/dL, segons el Mount Sinaii.
- Un valor d'albumina de 418.00 g/dL està molt per sobre del rang normal i és extremadament inusual.

https://www.mountsinai.org/health-library/tests/albumin-blood-serum-test

Cobre (4.00-588.00 µg/dL):

- El rang normal de coure total en la sang és de 70 a 140 µg/dL per a adults. Per tant, un valor de 588 µg/dL també està molt per sobre del rang normal i és inusual.
- Nivells tan alts podrien ser indicatius de condicions mèdiques específiques com la malaltia de Wilson, entre d'altres.

https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=167&ContentID=total_copper_blood

Alk_Phos (289.00-13862.40 U/liter):

- El rang normal de fosfatasa alcalina (Alk_Phos) en la sang és de 40 a 129 U/L, segons la Mayo Clinic.
- Un valor d'Alk_Phos de 13,862.40 U/L està molt per sobre d'aquest rang normal i és extremadament inusual. Nivells tan elevats podrien indicar condicions mèdiques serioses, com malalties hepàtiques o òssies, però són atípics.

https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595

SGOT (26.35-457.25 U/ml):

- El rang normal per a AST (SGOT) varia de 8 a 33 U/L, segons Cleveland Clinic.
- Un valor de SGOT de 457.25 U/L també està molt per sobre del rang normal i és inusual. Aquests nivells elevats podrien indicar dany hepàtic o muscular.

https://my.clevelandclinic.org/health/diagnostics/22147-aspartate-transferase-ast

Triglicèrids (33.00-598.00 mg/dL):

- El rang normal de triglicèrids en sang per a adults és inferior a 150 mg/dL. Un nivell superior a 500 mg/dL se considera molt alt. Per tant, un valor de 598 mg/dL està dins del rang que es considera molt alt, segons la Mayo Clinici Cleveland Clinic. Aquests nivells elevats poden incrementar el risc de malalties del cor i pancreatitis.

https://my.clevelandclinic.org/health/articles/11117-triglycerides

Plaquetes (62.00-721.00 x10^3/µL):

- Per a plaquetes, el rang normal varia segons la font, però generalment es troba entre 150,000 i 450,000 plaquetes per microlitre de sang. Un valor de 62,000 plaquetes per microlitre és considerat baix, ja que està per sota del rang normal, mentre que un valor de 721,000 plaquetes per microlitre és considerat elevat.

https://my.clevelandclinic.org/health/diagnostics/21782-platelet-count

Temps de Protrombina (1.02-416.00 segons):

- El rang normal de PT en la sang és d'aproximadament 10 a 13.5 segons, i l'INR normal és de 0.8 a 1.1 per a persones sanes. Un temps de protrombina (PT) més alt indica que el teu cos triga més temps del normal a formar coàguls. Els nivells de PT de 416.00 segons estan significativament per sobre del rang normal, la qual cosa podria indicar un problema amb el procés de coagulació de la sang.

https://my.clevelandclinic.org/health/diagnostics/17691-prothrombin-time-pt-test

Etapas (Stage) (1-4):

- En el context mèdic, l'etapa fa referència a l'etapa d'una malaltia o condició, com el càncer. Aquestes etapes són generalment classificades en un rang numèric més baix (per exemple, de l'etapa 1 a l'etapa 4 en el càncer).

https://www.cancer.gov/about-cancer/diagnosis-staging/staging

```
Python
Edat màxima en anys: 78.49 anys
Edat mínima en anys: 26.30 anys
```

Aquests valors són consistents amb els rangs d'edat normals per a adults. L'edat màxima de 78.49 anys és una expectativa de vida raonable, i l'edat mínima de 26.30 anys cau dins de l'etapa adulta.

```
Python
N_Days màxima en anys: 13.14 anys
N_Days mínima en anys: 0.11 anys
```

Aquests valors semblen indicar el temps transcorregut, possiblement des de l'inici d'un estudi o tractament fins a un esdeveniment específic (com la mort, el trasplantament o l'anàlisi final de l'estudi). El rang de 0.11 anys (aproximadament 40 dies) a 13.14 anys és plausible per a un seguiment en estudis longitudinals o clínics.

En resum, cap dels valors detectats dels possibles outliers, realment ho son, ja que realment poden succeir.

2.1.5 Feature engineering

2.1.5.1 Construction

Ascites, Hepatomegaly, i Spiders: Aquests atributs estan relacionats amb símptomes físics de la cirrosi hepàtica. Podrien ser combinats per crear un nou atribut que reflecteixi la gravetat general dels símptomes físics. Gràcies a l'estudi del "2.1.4 Detecció e imputació d'outliers", hem pogut treure aquesta relació.

2.1.5.2 Selection

N_Days: Aquest atribut representa el temps transcorregut fins a l'esdeveniment (mort, trasplantament, o anàlisi de l'estudi), el que pot estar correlacionat amb l'estatut final del pacient però no necessàriament prediu la supervivència futura.

Nota Addicional sobre la Variable "Age"

En el nostre conjunt de dades, la variable "Age" originalment estava representada en dies, la qual cosa resultava en un rang de valors molt alt i podia ser difícil d'interpretar. Per facilitar la comprensió i anàlisi d'aquesta variable, hem convertit l'edat de dies a anys.

Finalment, és important destacar que la 'Codificació de variables' es realitza al final de la secció de '3. Preparació de variables'. Aquesta estratègia s'adopta per evitar múltiples ramificacions en el conjunt de dades. Tenint en compte que alguns models accepten variables categòriques mentre que altres no, aquest enfocament permet mantenir una base de dades unificada sense dividir-la innecessàriament. Així, s'assegura una gestió més eficient de les dades i s'optimitza el procés de modelatge.

2.2 Gestió de Dades Post-Partició

2.2.1 Partició

En aquesta etapa del projecte, hem realitzat una partició de les dades després d'un procés de barreja aleatòria (shuffle). Aquesta barreja assegura que les dades es reparteixin de forma equitativa i que no hi hagi sesgos en la distribució dels conjunts de trenament i test. A continuació, hem procedit a la divisió de les dades en conjunts de trenament i test, optant per una sola partició, ja que implementarem validació creuada (cross-validation). Com s'ha observat en la secció "1.2 Primera visió", la validació creuada és una estratègia eficaç per als models que volem aplicar. A més, la nostra base de dades té un nombre relativament petit d'individus, i l'ús de la validació creuada permetrà aprofitar millor aquestes dades limitades, incrementant la fiabilitat i la robustesa dels nostres models predictius.

2.2.2 Tractament de missings

Per abordar aquest repte, hem implementat una estratègia en dues fases: primer, determinar el millor mètode d'imputació per a les nostres dades; segon, aplicar aquest mètode a la base de dades original X_train.

2.2.2.1 Determinació del Millor Mètode d'Imputació

Per identificar el mètode d'imputació més adequat, vam crear un conjunt de dades amb missings autoimputats. Aquest enfocament ens va permetre simular un escenari realista de dades incompletes i avaluat la efectivitat de diferents tècniques d'imputació tant per a variables numèriques com categòriques.

Variables Numèriques

Per a les variables numèriques, es van examinar dos mètodes coneguts per la seva eficàcia:

- **Random Forest:** Un algoritme basat en arbres que pot manejar de manera eficaç les complexitats en les dades numèriques.

- **MICE (Multiple Imputation by Chained Equations):** Un mètode sofisticat que realitza múltiples imputacions considerant les interrelacions entre variables.

Variables Categòriques

Per a les variables categòriques, vam provar els següents mètodes:

- **Moda:** La imputació per la moda és una tècnica senzilla on es substitueixen els valors perduts per la categoria més freqüent.
- **Hot-Deck:** Aquest mètode implica emplenar els valors perduts amb valors observats en casos similars dins del mateix conjunt de dades.

2.3.2.2 Resultats de l'Avaluació

Després d'aplicar aquests mètodes a la nostra base de dades amb missings autoimputats, vam realitzar una anàlisi comparativa per determinar quins d'ells oferien els millors resultats. Per a les dades numèriques, el mètode Random Forest va demostrar ser el més eficaç, mentre que per a les dades categòriques, la tècnica de la Moda va resultar ser la més adequada.

Amb aquests resultats, vam procedir a la fase final del nostre tractament de missings. Vam aplicar l'imputació mitjançant Random Forest per a les variables numèriques i Moda per a les variables categòriques en la nostra base de dades original X_train.

```
Python
Missings després del preprocessament en X_test:
Age                          0
Bilirubin                    0
Cholesterol                  0
Albumin                      0
Copper                       0
Alk_Phos                     0
SGOT                         0
Tryglicerides                0
Platelets                    0
```

```
Prothrombin                0
Stage                      0
Physical_Symptoms_Score    0
Drug                       0
Sex                        0
Edema                      0
```

```python
Missings després del preprocessament en X_train:
Age                        0
Bilirubin                  0
Cholesterol                0
Albumin                    0
Copper                     0
Alk_Phos                   0
SGOT                       0
Trygliceride               0
Platelets                  0
Prothrombin                0
Stage                      0
Physical_Symptoms_Score    0
Drug                       0
Sex                        0
Edema                      0
```

2.2.3 Tractament de balanceig

2.2.3.1 Anàlisi dels Mètodes de Balanceig

RandomUnderSampler i RandomOverSampler: Aquests mètodes podrien no ser els més adequats en aquest cas. El submostreig (under-sampling) de les categories més freqüents (C i D) podria resultar en la pèrdua d'informació important, mentre que l'augment (over-sampling) de la categoria CL podria conduir a un sobreajustament degut a la repetició excessiva de les poques mostres existents.

TomekLinks i ClusterCentroids: Aquests mètodes estan més enfocats en millorar les fronteres entre les classes que en equilibrar les seves distribucions. Potser no siguin suficients per a abordar el desequilibri significatiu present en les dades.

SMOTE i SMOTETomek: Aquests mètodes podrien ser més adequats. SMOTE crea mostres sintètiques de la classe minoritària (en aquest cas, CL), millorant l'equilibri sense perdre informació important. SMOTETomek combina SMOTE amb el submostreig utilitzant Tomek Links per a millorar encara més la qualitat de les dades balancejades.

2.2.3.2 Conclusió

Tenint en compte l'objectiu de l'estudi i la distribució de la variable objectiu, utilitzarem SMOTE o SMOTETomek. Aquests mètodes poden augmentar efectivament la representació de la categoria minoritària (CL) sense perdre informació crítica de les altres categories. A més, la creació de mostres sintètiques pot ajudar a evitar el sobreajustament, a la vegada que millora la capacitat del model per aprendre sobre la categoria menys representada.

I finalment a quedat:

```
Python
Distribució de classes abans de SMOTETomek: Counter({'C': 183, 'D': 128, 'CL': 23})
Distribució de classes després de SMOTETomek: Counter({'CL': 175, 'D': 165, 'C': 165})
```

2.2.4. Codificació de variables

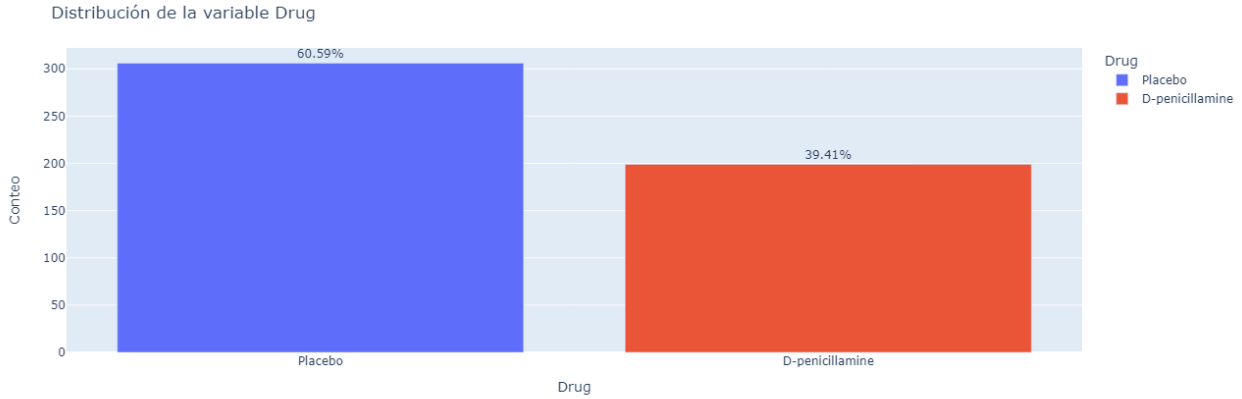
Com es va observar a la secció "1.2 Primera visió", models com KNN i SVM requereixen dades numèriques, per la qual cosa és necessari transformar les variables categòriques en numèriques. La nostra base de dades conté quatre variables categòriques: "Status", "Drug", "Sex" i "Edema". "Status" és la variable objectiu i per tant no requerirà transformació.

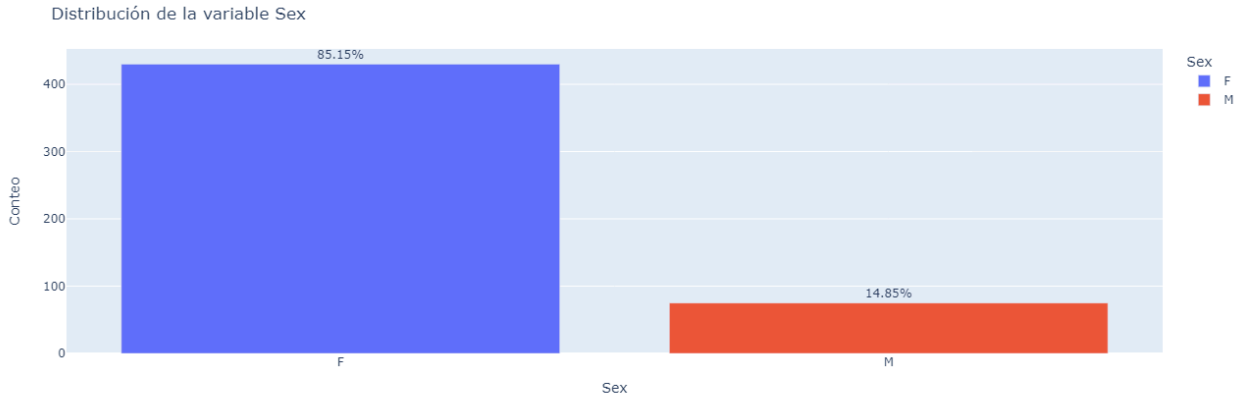
Per les altres tres variables ("Drug", "Sex", "Edema"), que són categòriques nominals (on l'ordre no importa), aplicarem la tècnica de codificació one-hot en lloc de label encoding. Aquesta decisió es basa en el fet que la codificació one-hot és més

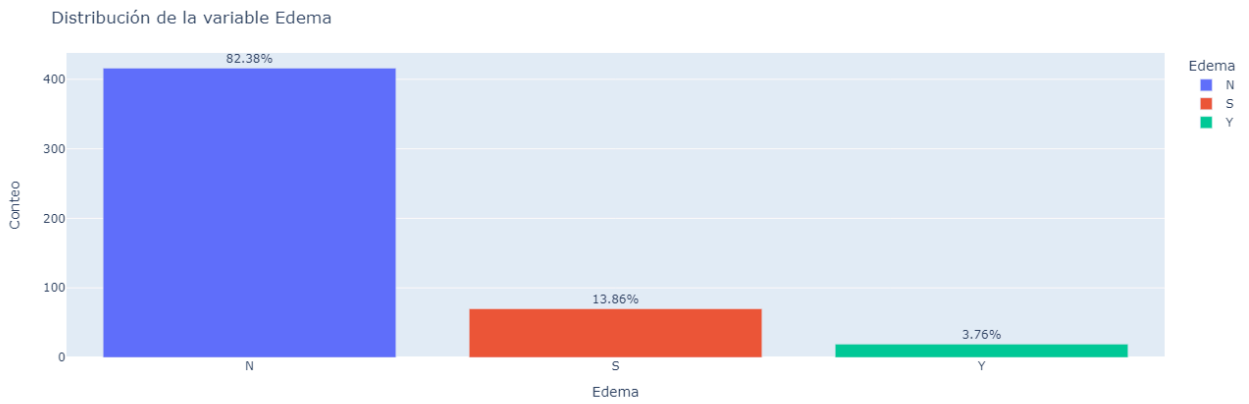
adequada per a aquest tipus de dades, ja que manté la distància equidistant entre les categories, evitant qualsevol suposició d'ordre o jerarquia entre elles, el que és especialment important per a models com SVM i KNN, que depenen de la distància entre les dades.

2.3.5 Análisis post processament

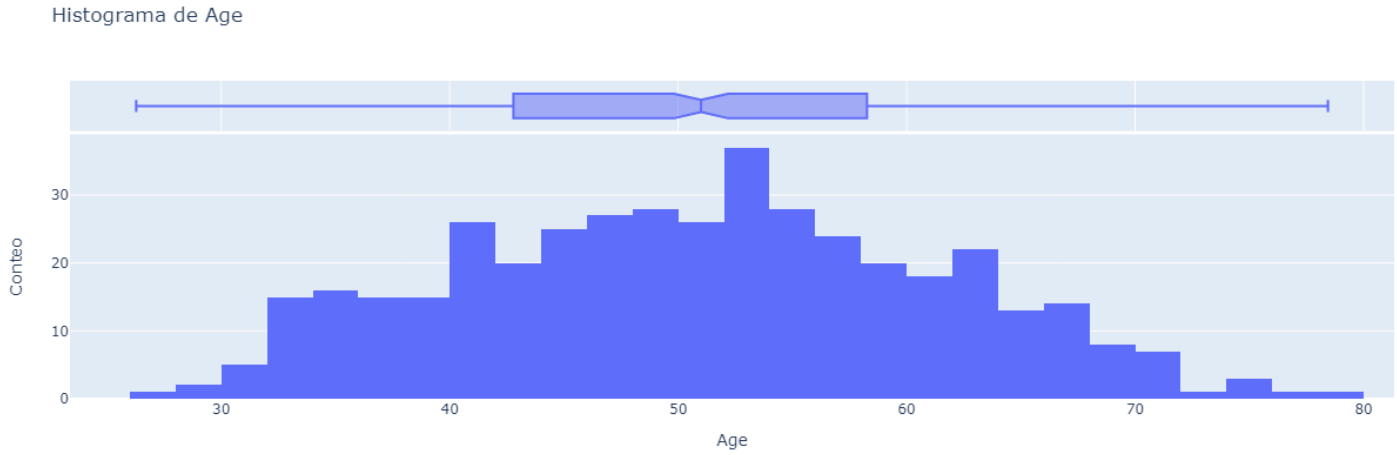
En aquesta secció, repetirem l'anàlisi univariant realitzat a la secció "2.1.1 Anàlisi univariant", però aquesta vegada amb les dades preprocessades. Utilitzarem codi similar per generar gràfiques que ens ajudin a visualitzar les distribucions de les variables categòriques i numèriques després del preprocessament.

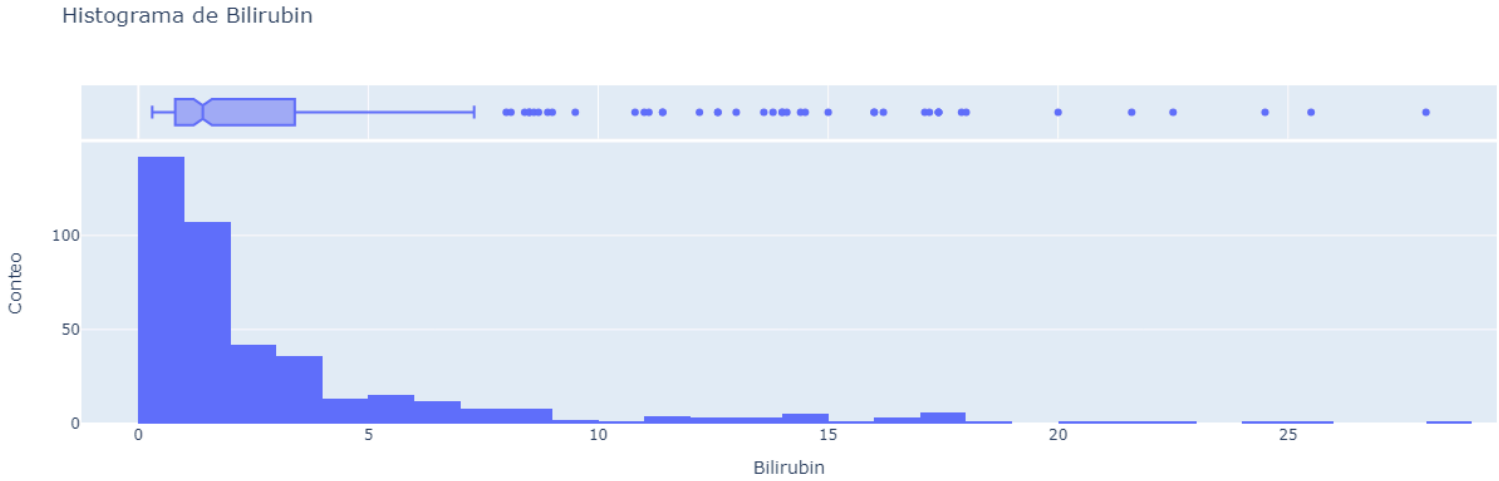


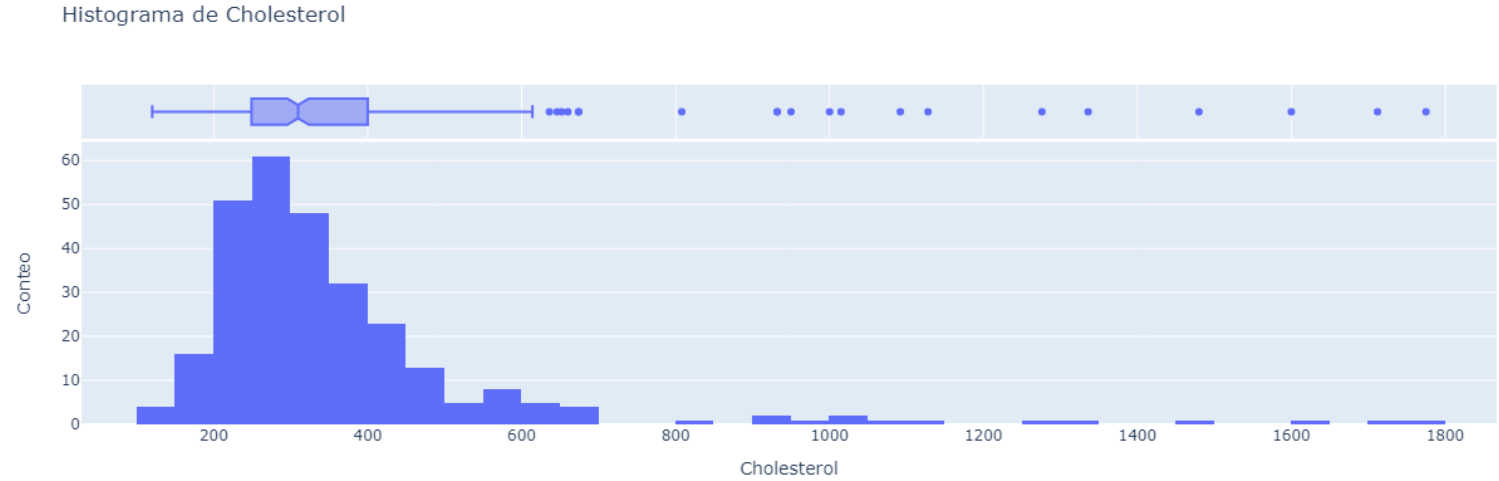




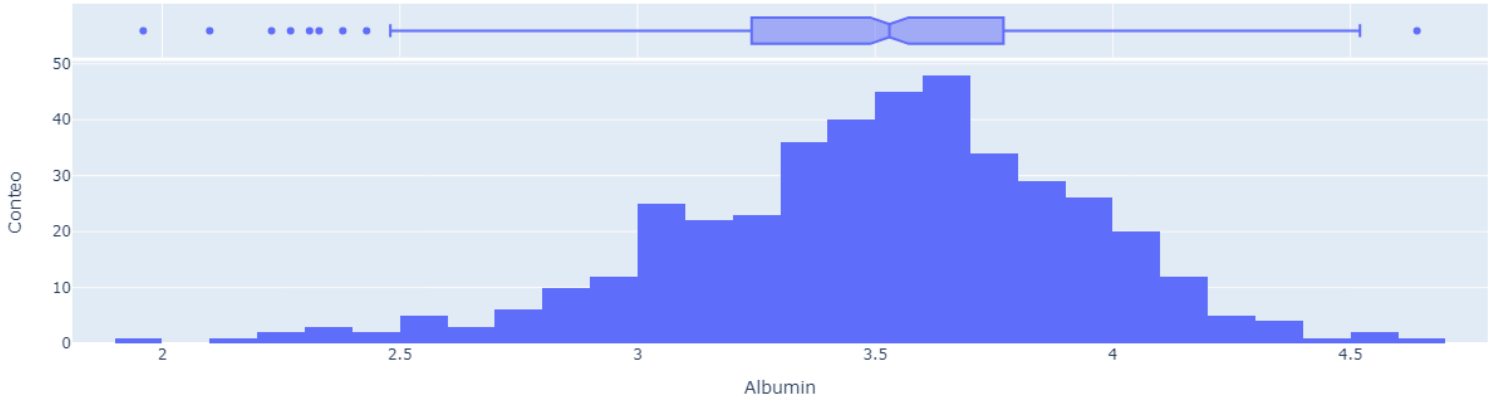
Les gràfiques mostren que els percentatges de les diferents categories en les variables preprocessades s'han mantingut similars a les originals. Aquesta consistència indica que el preprocessament ha estat efectiu i no ha alterat significativament la distribució d'aquestes variables.



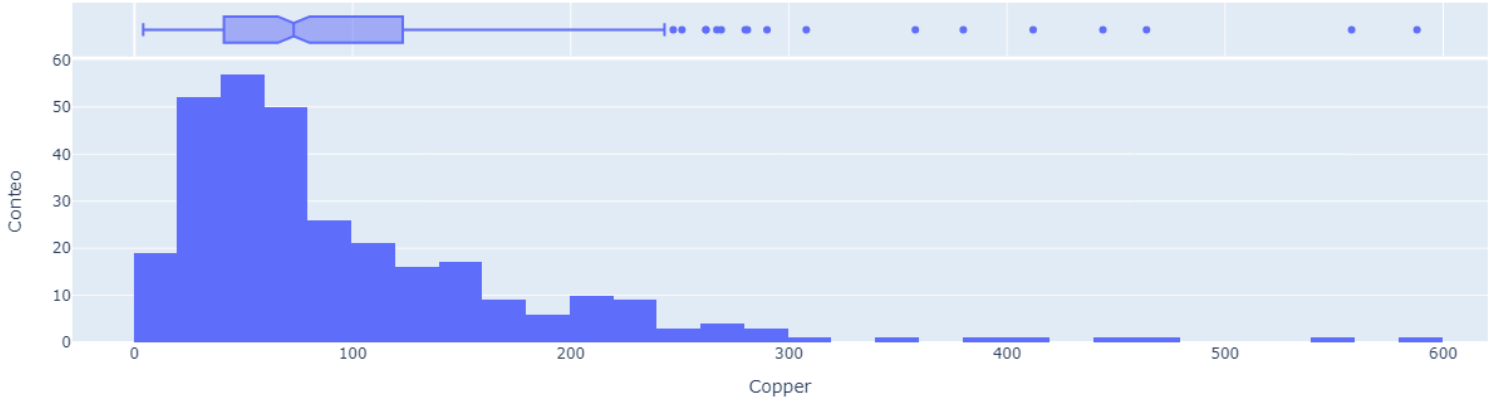

Histograma de Bilirubin
Conteo
100
50
0
0
5
10
15
20
25
Bilirubin


Histograma de Cholesterol
Conteo
60
50
40
30
20
10
0
200
400
600
800
1000
1200
1400
1600
1800
Cholesterol

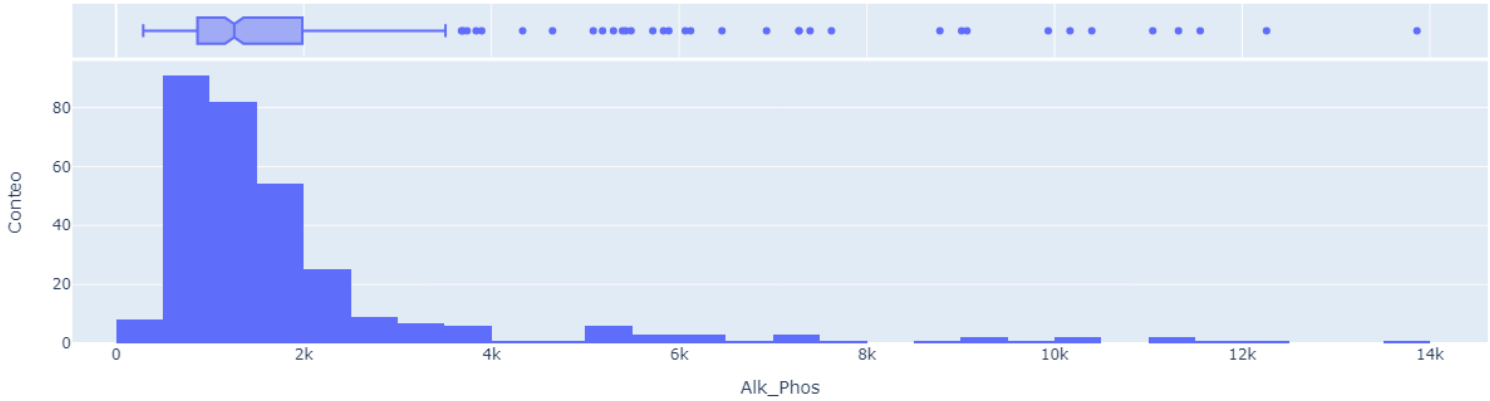
Histograma de Albumin



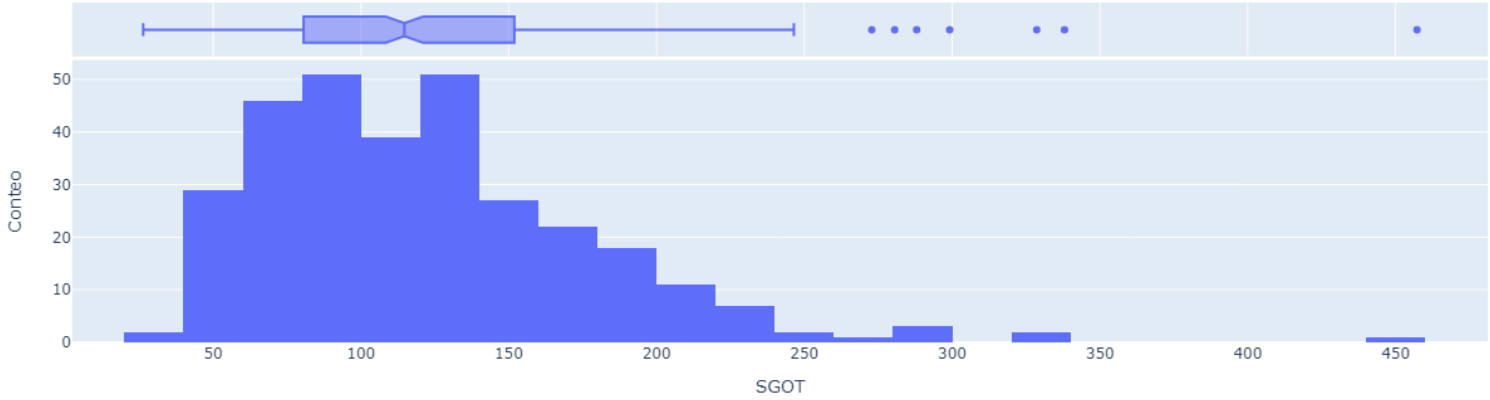
Histograma de Copper

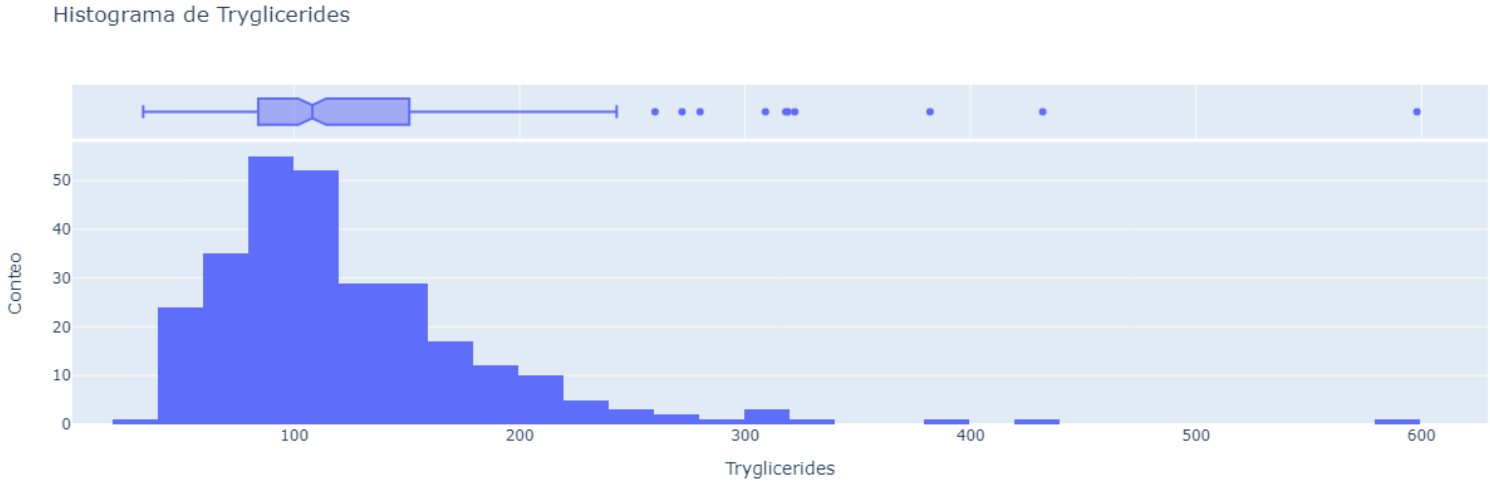


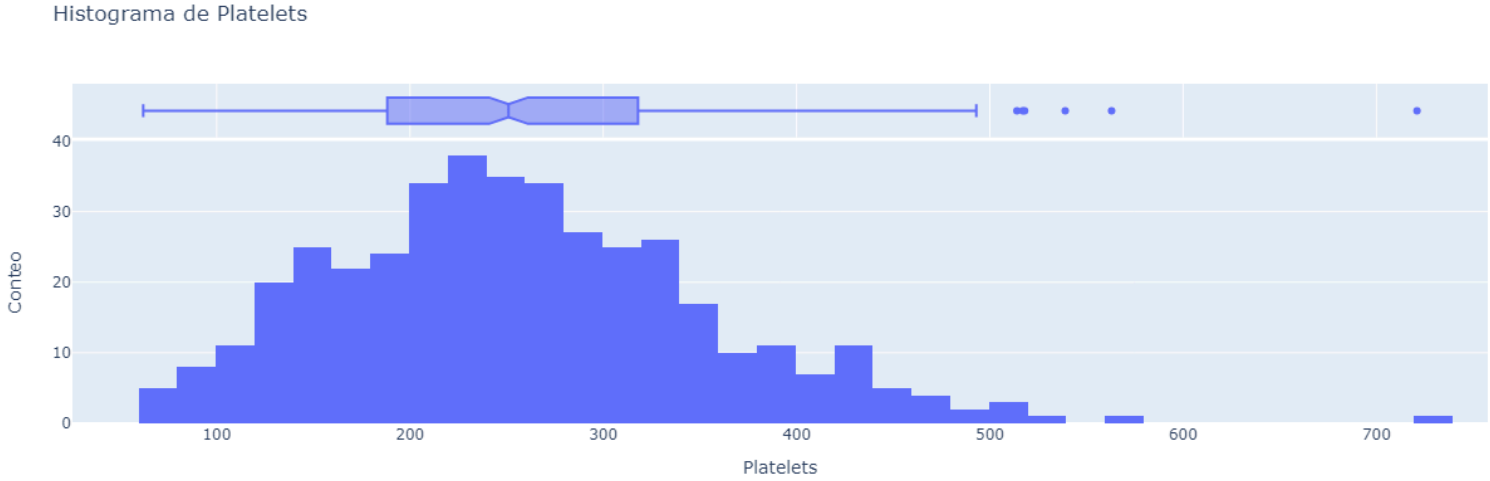
Histograma de Alk_Phos



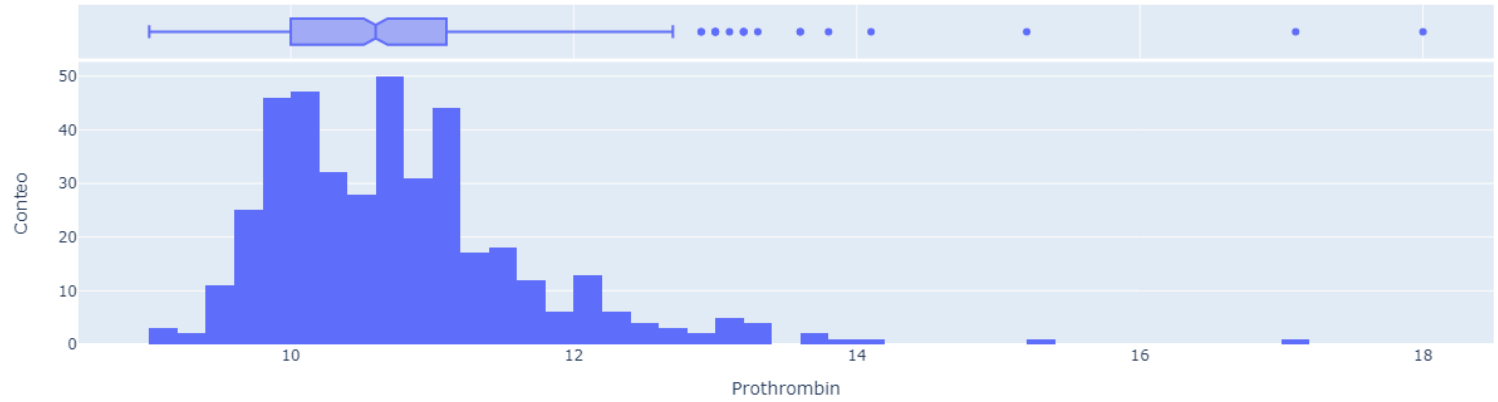
Histograma de SGOT



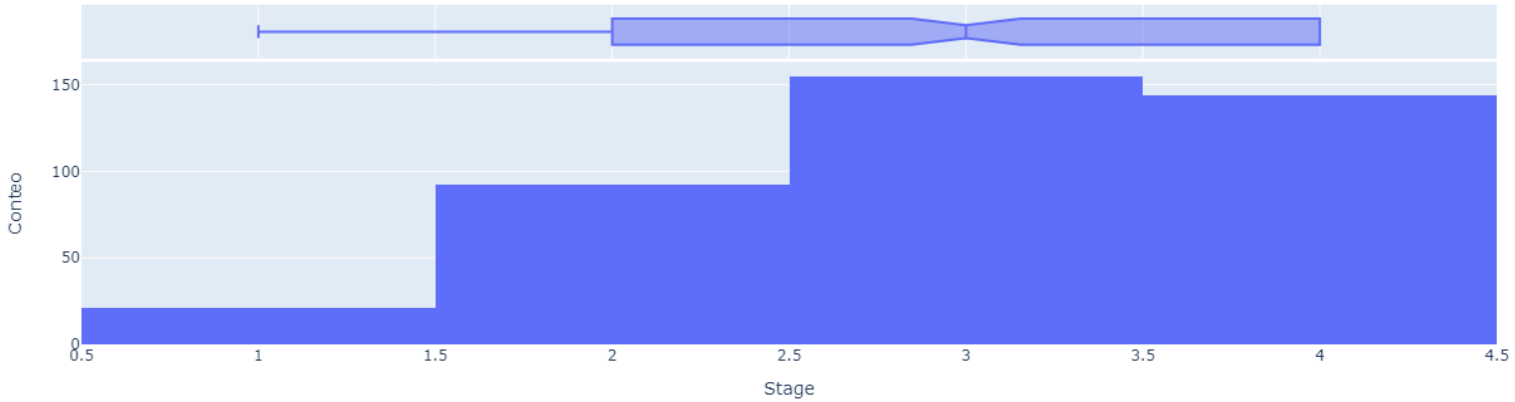

Histograma de Trygliceridos
Conteo
0
10
20
30
40
50
100
200
300
400
500
600
Trygliceridos

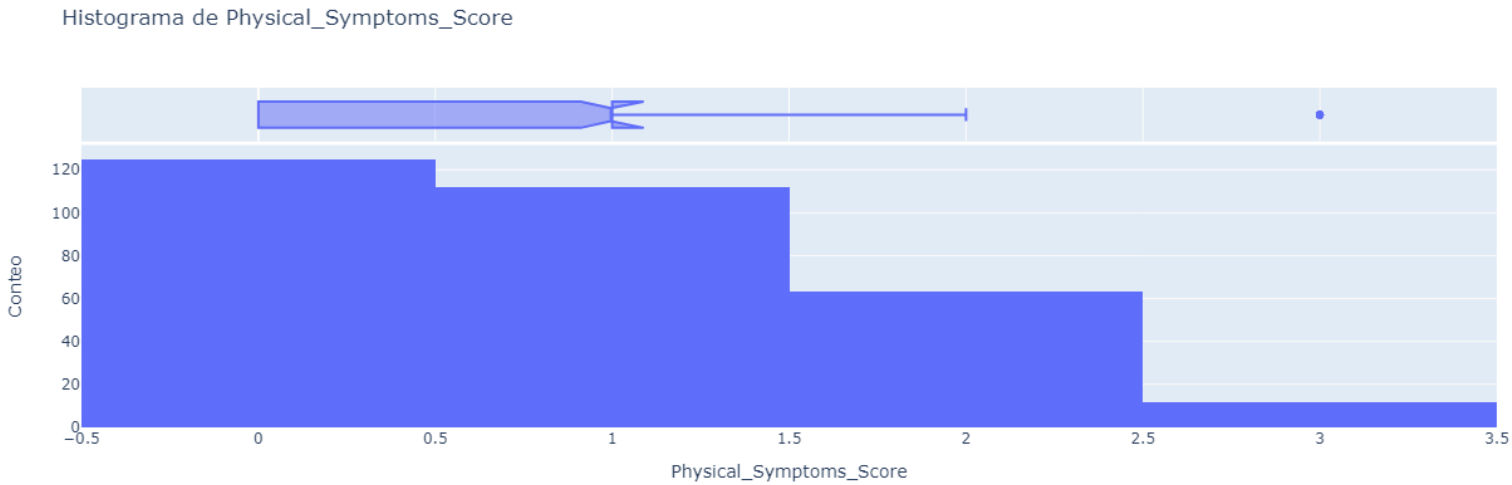

Histograma de Platelets
Conteo
0
10
20
30
40
100
200
300
400
500
600
700
Platelets

Histograma de Prothrombin



Histograma de Stage





L'anàlisi a través d'histogrames i diagrames de caixa mostra que la distribució de les variables numèriques també roman essencialment inalterada. Els punts observats en els diagrames de caixa, que abans podrien haver semblat outliers, ara ja sabem que realment no ho son.

3. PREPARACIÓ DE VARIABLES

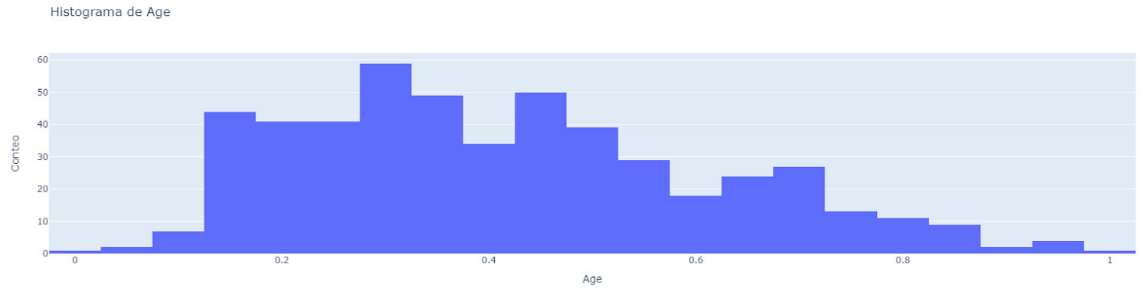
En aquesta secció, l'anàlisi es centrarà en la base de dades numèriques que han estat preprocessades mitjançant la tècnica de one-hot encoding. Aquest enfocament es pren per evitar la repetició del análisis.

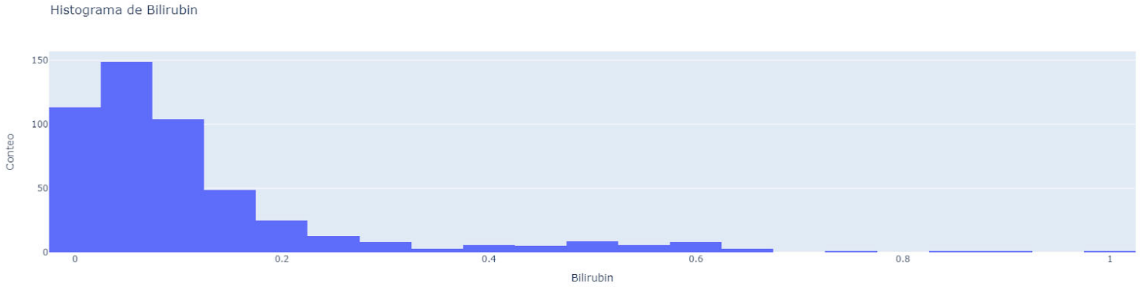
3.1 Normalització i escalat

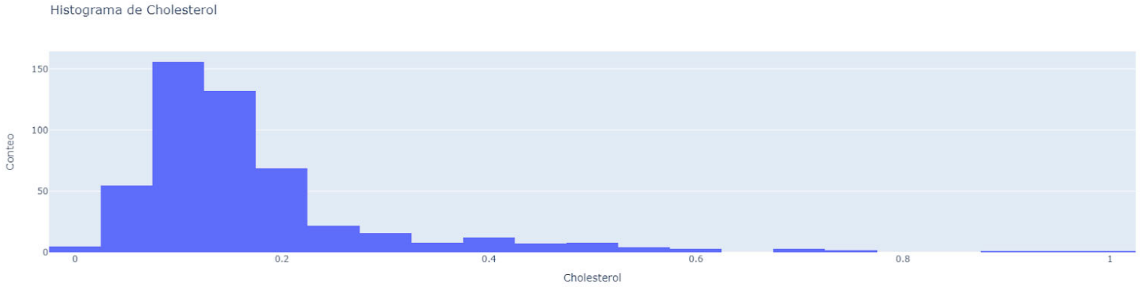
Com s'ha observat en la secció "1.2 Primera Visió", la normalització i l'estandardització de les variables poden millorar significativament l'eficàcia dels models de machine learning. Durant la secció "2. Anàlisi i Preprocessat de Dades", hem identificat que les nostres variables encara no han estat normalitzades ni estandarditzades. Per tant, en aquesta etapa, procedirem a aplicar aquestes tècniques.

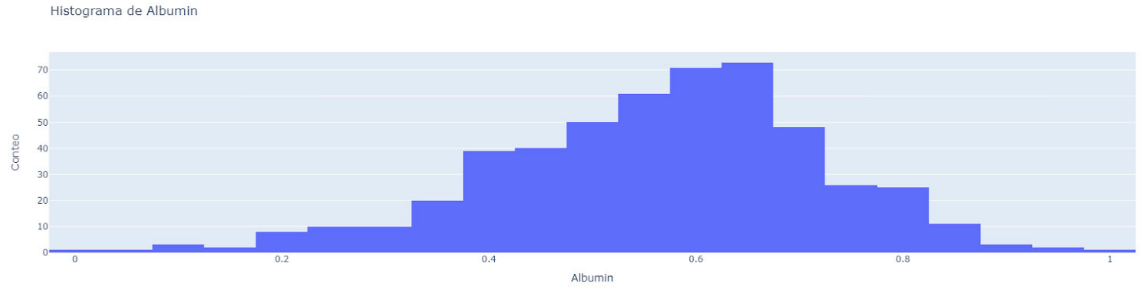
Per a la normalització, s'ha utilitzat MinMaxScaler. Aquesta tècnica ajusta les dades de manera que totes les característiques es trobin dins d'un rang específic, habitualment entre 0 i 1. Aquest mètode és particularment útil quan tenim característiques amb diferents escales i volem homogeneïtzar aquestes en un mateix rang.

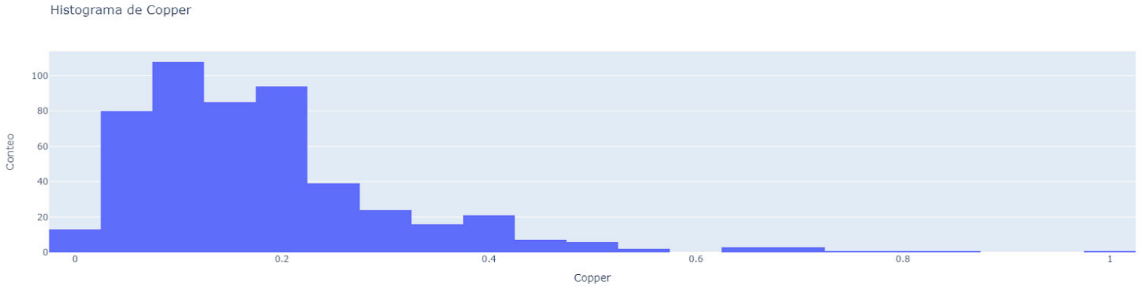
D'altra banda, per a l'estandardització, s'ha aplicat StandardScaler. Aquesta aproximació transforma les dades perquè tinguin una mitjana de zero i una desviació estàndard d'una unitat.

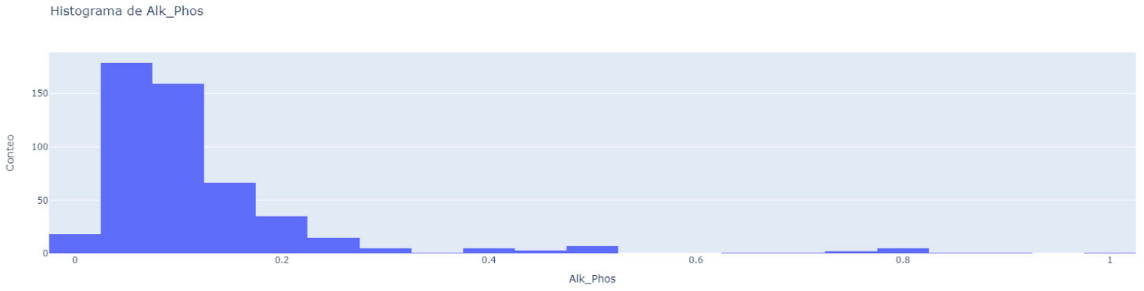

Histograma de Age
Conteo
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
Age

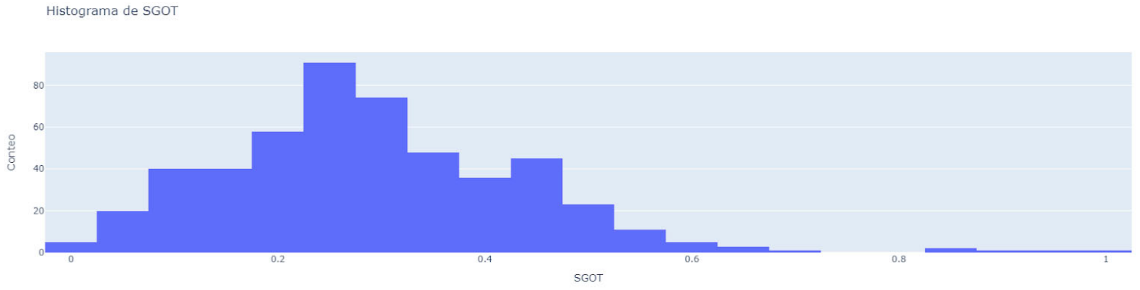

Histograma de Bilirubin
Conteo
150
100
50
0
0
0.2
0.4
0.6
0.8
1
Bilirubin

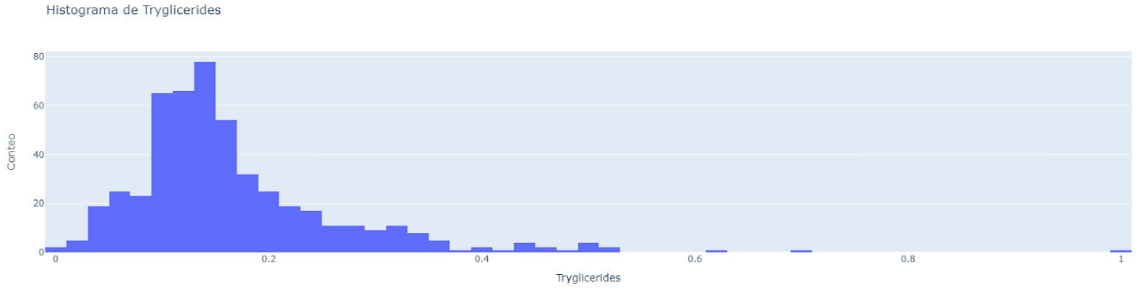

Histograma de Cholesterol
Conteo
150
100
50
0
0
0.2
0.4
0.6
0.8
1
Cholesterol

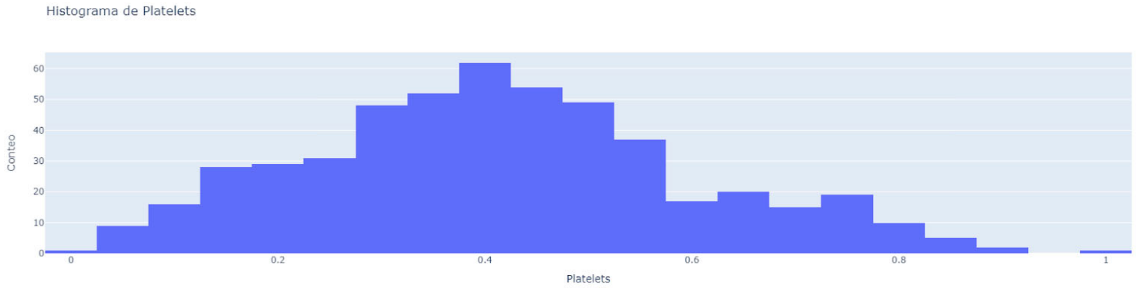

Histograma de Albumin
Conteo
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
Albumin

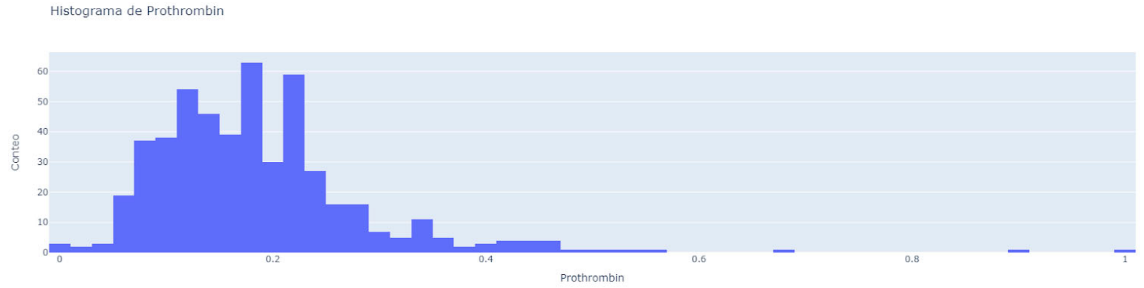

Histograma de Copper
Conteo
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
Copper

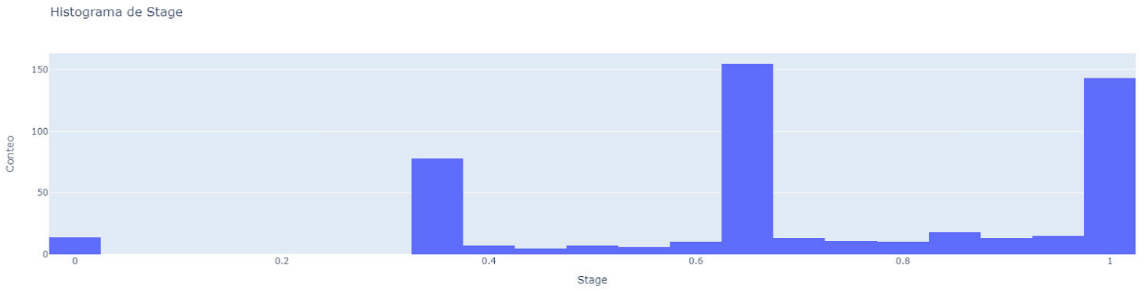

Histograma de Alk_Phos
Conteo
150
100
50
0
0
0.2
0.4
0.6
0.8
1
Alk_Phos

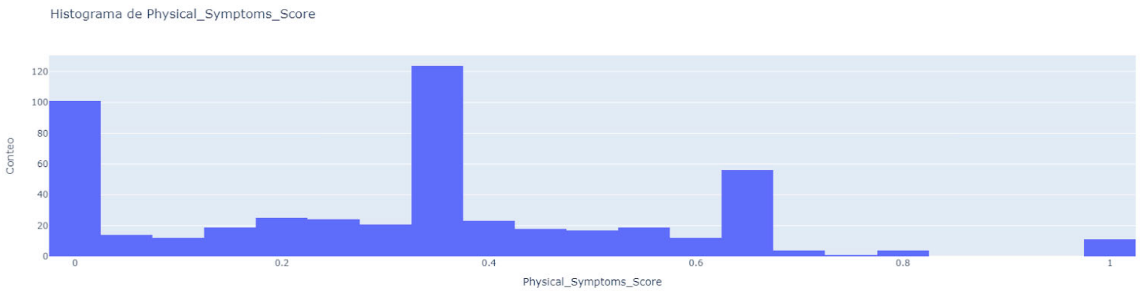

Histograma de SGOT
Conteo
0
20
40
60
80
0
0.2
0.4
0.6
0.8
1
SGOT


Histograma de Tryglicerides
Conteo
0
20
40
60
80
0
0.2
0.4
0.6
0.8
1
Tryglicerides


Histograma de Platelets
Conteo
0
10
20
30
40
50
60
0
0.2
0.4
0.6
0.8
1
Platelets


Histograma de Prothrombin
Conteo
Prothrombin
0
10
20
30
40
50
60
0
0.2
0.4
0.6
0.8
1


Histograma de Stage
Conteo
Stage
0
50
100
150
0
0.2
0.4
0.6
0.8
1


Histograma de Physical_Symptoms_Score
Conteo
Physical_Symptoms_Score
0
20
40
60
80
100
120
0
0.2
0.4
0.6
0.8
1

3.2 Correlacions entre variables numèriques

En el marc del desenvolupament d'un model predictiu per a pacients amb cirrosi hepàtica, és imperatiu comprendre les interaccions entre les variables clíniques. Aquesta comprensió es pot afavorir mitjançant un anàlisi de correlació, especialment quan s'utilitzen models com KNN i SVM, que són particularment sensibles a la colinealitat i a les diferents escales de variables.

Per abordar aquesta qüestió, hem triat la prova ANOVA, la qual mesura la significació estadística de la diferència entre les mitjanes de grups. En el nostre cas no podem aplicar una matriu de correlació ja que la nostre variable objectiu és una categórica.

```
Python
Variable: Age, P-Value: 0.000000000, Significativa: Sí
Variable: Bilirubin, P-Value: 0.000000000, Significativa: Sí
Variable: Cholesterol, P-Value: 0.000000001, Significativa: Sí
Variable: Albumin, P-Value: 0.000000000, Significativa: Sí
Variable: Copper, P-Value: 0.000000000, Significativa: Sí
Variable: Alk_Phos, P-Value: 0.000020885, Significativa: Sí
Variable: SGOT, P-Value: 0.000000000, Significativa: Sí
Variable: Tryglicerides, P-Value: 0.000071996, Significativa: Sí
Variable: Platelets, P-Value: 0.000000001, Significativa: Sí
Variable: Prothrombin, P-Value: 0.000000000, Significativa: Sí
Variable: Stage, P-Value: 0.000000000, Significativa: Sí
Variable: Physical_Symptoms_Score, P-Value: 0.000000000, Significativa: Sí
```

Els resultats mostren que totes les variables numèriques estudiades presenten una correlació significativa amb el 'Status' dels pacients. Els valors P de l'ANOVA són considerablement baixos, ben per sota del llindar de 0.05, indicant que cada variable numèrica té una associació estadísticament significativa amb l'evolució de la cirrosi hepàtica en pacients.

3.3 Anàlisi de variables categòriques i variable objectiu

Per a l'anàlisi de les relacions entre les variables categòriques i la variable objectiu en aquesta fase de la recerca, hem aplicat el test de Chi-quadrat. Aquest test és una tècnica estadística que s'utilitza per determinar si hi ha una associació significativa entre dues variables categòriques. És particularment útil en aquest context, ja que tant les variables independents com la variable objectiu ('Status') són categòriques.

3.3.1 Interpretació dels Resultats

```
Python
Variable: Drug, P-Value: 0.35691, Significativa: No
Variable: Sex, P-Value: 0.00760, Significativa: Sí
Variable: Edema, P-Value: 0.00000, Significativa: Sí
```

Variable 'Drug': Amb un P-Value de 0.35691, aquest resultat indica que no hi ha una associació estadísticament significativa entre el tipus de medicament i la supervivència dels pacients. Això suggerix que la variable 'Drug' podria no ser un predictor útil per a la supervivència dels pacients amb cirrosi hepàtica en el nostre model.

Variable 'Sex': El P-Value obtingut és de 0.00760, el que indica una associació significativa entre el gènere del pacient i la seva supervivència. Això podria suggerir que el gènere té un paper important en la predicció dels resultats dels pacients.

Variable 'Edema': Amb un P-Value de 0.00000, aquest resultat mostra una associació molt forta i estadísticament significativa entre la presència d'edema i la supervivència dels pacients. Això indica que l'edema és un factor rellevant a considerar en el model predictiu.

3.4 Eliminació de variables redundants o sorolloses

En el desenvolupament del model predictiu per a pacients amb cirrosi hepàtica, una etapa crítica és l'eliminació de variables que no contribueixen de manera significativa a la capacitat predictiva del model. Aquesta selecció s'ha basat en els resultats obtinguts en les seccions "3.2 Correlacions entre variables numèriques" i "3.3 Anàlisi de variables categòriques i variable objectiu".

A través de l'aplicació d'una prova ANOVA per a les variables numèriques i el test de Chi-quadrat per a les variables categòriques, hem determinat que la variable 'Drug' no mostra una associació estadísticament significativa amb la variable objectiu, 'Status'. Això indica que 'Drug' no és un predictor útil per a la supervivència dels pacients en aquest context específic.

Paral·lelament, podem extreure que introduir **pracebo** a un pacient no altera ni millora les conseqüències del pacient.

Per tant, per al conjunt de dades X_train_balanced, que serà utilitzat en l'aplicació del model de Random Forest —un model que pot manejar variables categòriques—, s'ha decidit eliminar la variable 'Drug'. Això ajudarà a simplificar el model sense sacrificar la seva capacitat predictiva.

Pel que fa a X_train_norm_est, que està preparat per a l'ús en models que requereixen codificació de tipus one-hot com SVM i KNN, s'eliminaren les variables 'Drug_Placebo' i 'Drug_D-penicillamine'. Aquestes variables són representacions binàries de la variable categòrica 'Drug', i la seva eliminació es basa en la mateixa raó: la falta d'associació significativa amb l'objectiu de la supervivència dels pacients.

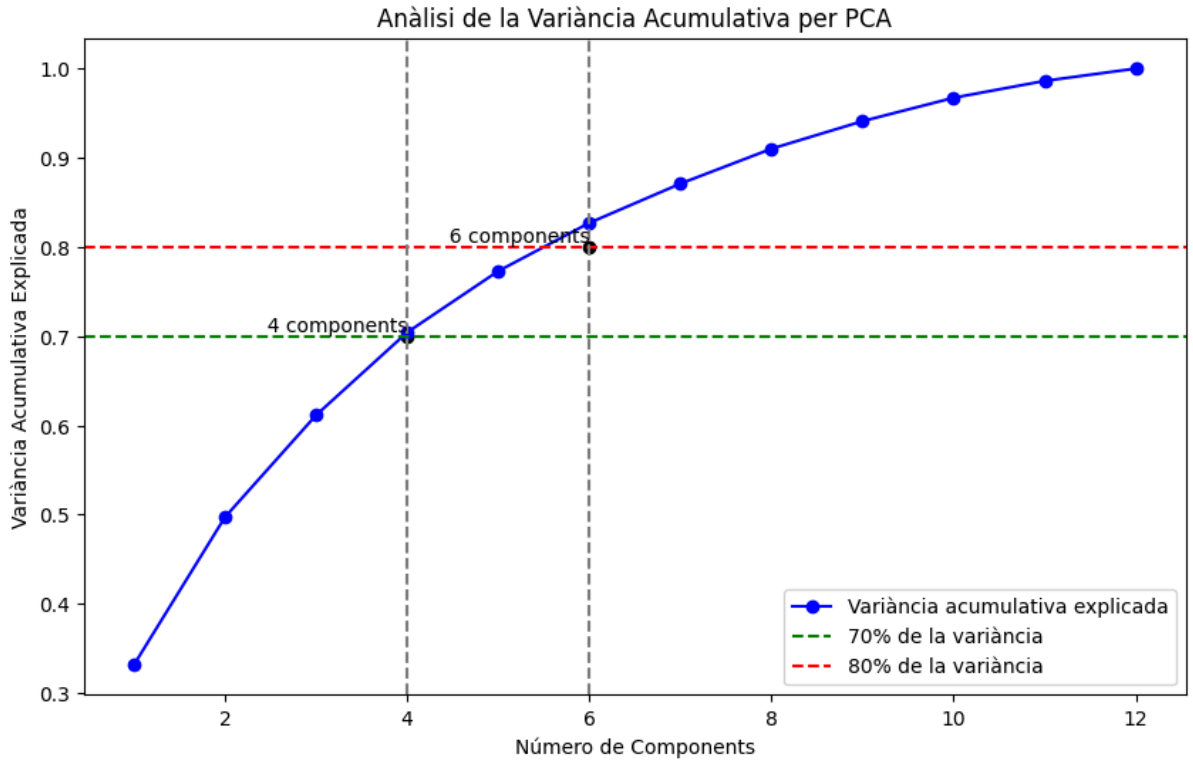
El passaria el mateix amb les dades de test.

3.5 Estudi de dimensionalitat amb PCA.

L'Anàlisi de Components Principals (PCA) és una eina estadística poderosa utilitzada per simplificar la complexitat en conjunts de dades de grans dimensions mitjançant

la reducció del nombre de variables.L'aplicació de PCA podria semblar beneficiosa per identificar les combinacions de característiques que més influencien en la variància dels dades, així com per millorar l'eficiència computacional i potencialment augmentar el rendiment del model.

3.5.1 Explicació dels resultats



Superació de Llindars de Variància: Segons els resultats de PCA, els primers quatre components expliquen més del 70% de la variància total. Aquesta és una indicació que aquestes dimensions podrien ser suficients per capturar la majoria de la informació rellevant. No obstant això, per assolir una comprensió més profunda, és necessari considerar fins a sis components per explicar més del 80% de la variància.

Punt d'Inflexió: L'observació d'un colze en la gràfica de Scree després del quart component suggereix que els components addicionals aporten menys informació incremental. Aquest fet podria ser un indicatiu per optar per un model més simplificat.

Decisió sobre Components a Utilitzar: Basant-nos en la variància acumulativa explicada, una selecció de quatre a sis components podria ser considerada per la creació del model reduït, depenent de l'equilibri entre la simplificació desitjada i la preservació de la informació.

3.5.2 Justificació de no aplicar la reducció amb PCA

Necessitat d'Explicabilitat: En la medicina, els models predictius sovint han de ser explicables als professionals de la salut i als pacients. Si un model es basa en components principals en lloc de variables clíniques reals, això pot fer que sigui més difícil justificar decisions clíniques basades en les prediccions del model.

Interaccions entre Característiques: Les malalties complexes com la cirrosi hepàtica poden tenir patrons de dades que són altament no lineals i que podrien ser capturats millor utilitzant les característiques originals en lloc de components principals que simplifiquen aquestes interaccions.

Complexitat vs. Benefici: La gràfica de variància acumulativa no mostra un punt on la reducció de dimensions addicionals deixi de ser beneficiosa significativament, lo qual implica que la complexitat inherent al conjunt de dades podria ser necessària per a la precisió del model.

Rendiment del Model: Sense una justificació forta basada en una millora del rendiment del model, l'aplicació de tècniques de reducció de dimensions com PCA podria no ser aconsellable. La validació i comparació del rendiment del model amb i sense PCA podrien proporcionar dades addicionals per a prendre aquesta decisió.

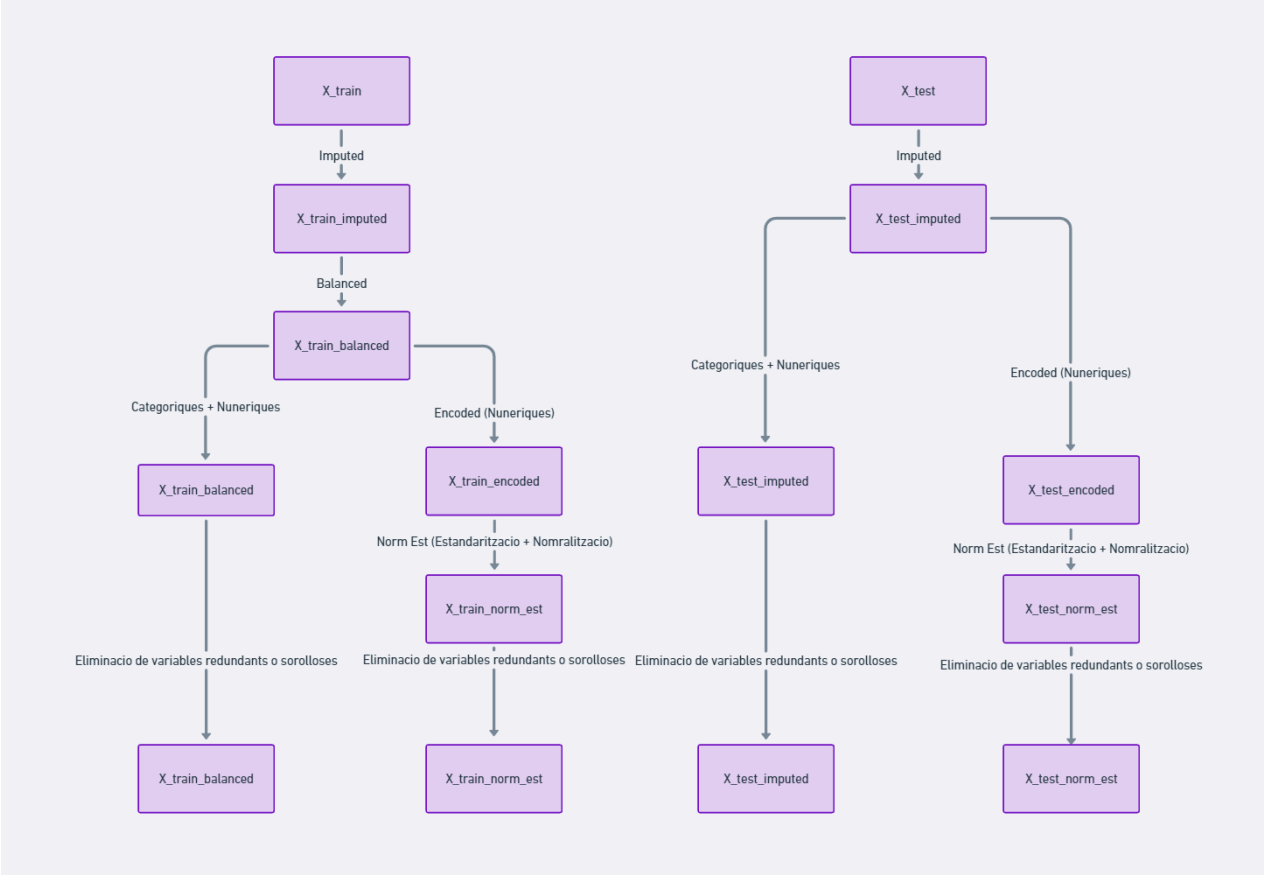
Tenint en compte aquests factors, es conclou que la reducció de dimensions mitjançant PCA no és aconsellable per aquest estudi específic. El manteniment de totes les variables clíniques originals, amb la seva plena complexitat i significació, és essencial per a la creació d'un model predictiu robust i interpretable per a la supervivència de pacients amb cirrosi hepàtica.

3.6 Ultimes consideracions

En aquest punt m'he adonat que la variable "Sex" és binària, per tant, per la base "X_train_norm_est", Sex_F i Sex_M és inversament proporcional, per la qual cosa es podria eliminar Sex_M. Per l'altra banda, per millorar l'eficàcia del model, per la base de dades "X_train_balanced" s'ha passat la variable "Sex" a numèrica, female=0 male=1.

3.7 Resum de les dedas

El procés de preparació i optimització de les dades és un component fonamental en el desenvolupament de models predictius robusts. A continuació, detallem el flux de treball seguit per la creació i evolució de les bases de dades utilitzades en aquest estudi, com es mostra en l'esquema proporcionat.



Així doncs, X_train_balanced i X_test_imputed s'utilitzaran per alimentar el model de l'Arbre de Decisions, assegurant que el tractament de les dades s'ajusta a les necessitats d'aquesta tècnica en particular. D'altra banda, X_train_norm_est i

X_test_norm_est estan preparats per ser utilitzats en els models SVM i KNN, proporcionant un terreny comú per a la comparació directa del rendiment d'aquests dos enfocaments. Aquesta estructuració final de les dades està en consonància amb la "1.2 Primera visió" i estableix una base sòlida per a la fase d'entrenament i validació dels models predictius. Ambdues utilitzen, y_train_balanced.

4. Definició de models

4.1 Definició de mètriques

Per determinar quina mètrica és la millor per avaluar els models en el context de la predicció de la supervivència de pacients amb cirrosi hepàtica, hem de considerar la naturalesa del problema i les implicacions de cada mètrica.

Recall (Sensibilitat): Aquesta mètrica mesura la proporció de positius reals (en aquest cas, pacients que no sobreviuen) que el model ha predit correctament. En un context mèdic on es prediu la supervivència, un alt recall és crucial perquè els casos de no supervivència són crítics i no haurien de ser passats per alt.

- **Raonament:** Si un pacient amb risc alt de no supervivència és incorrectament classificat com a baix risc, les conseqüències podrien ser greus.

Accuracy: Aquesta mètrica mesura la proporció de totes les prediccions correctes (tant positives com negatives) entre totes les prediccions realitzades. Malgrat que proporciona una visió general del rendiment del model, pot ser enganyosa en conjunts de dades desequilibrats, on una classe és molt més freqüent que l'altra.

- **Raonament:** La precisió pot ser alta simplement perquè el model prediu sempre la classe majoritària, però això no seria útil en un context mèdic on s'han d'identificar correctament tots els casos crítics.

F1 Score: És la mitjana harmònica de la precisió i el recall. Aquesta mètrica equilibra el recall i la precisió i és particularment útil quan les classes són desequilibrades, que és comú en conjunts de dades mèdiques.

- **Raonament:** F1 és útil quan volem un equilibri entre la identificació de tots els casos crítics (alta sensibilitat) i la minimització de les falses alarmes (alta precisió).

Tenint en compte aquestes consideracions:

- Per la gravetat de passar per alt un pacient que realment està en risc, el **Recall** és essencial.
- Però, com que també és important no sobrecarregar els recursos mèdics amb falsos positius, **l'F1 Score** proporciona un equilibri raonable entre el recall i la precisió.

Per tant, **l'F1 Score** seria probablement la millor mètrica per a aquesta aplicació específica, ja que busca un equilibri entre la identificació fiable de tots els pacients en risc (alt recall) i la minimización del nombre de falsos positius (alta precisió), cosa que és crucial en un entorn mèdic on cada decisió pot tenir conseqüències significatives.

4.2 Primer mode (KNN)

4.2.1 Motivació

Dins dels models considerats —KNN, Arbre de Decisions, i SVM—, KNN es destaca com una opció prometedora per diverses raons, explicades en "1.1.1 KNN".

4.2.2 Discussió dels hiperparàmetres disponibles, i dels valors usats

Per desenvolupar un model KNN per a la predicció de supervivència de pacients amb cirrosi hepàtica, és crucial seleccionar els hiperparàmetres adequats. Utilitzant la tècnica de Grid Search, explorem diferents combinacions d'hiperparàmetres per

optimitzar el model. Aquesta tecnica és particularment útil en aquest cas, ja que no teníem un coneixement previ clar sobre quins serien els millors hiperparàmetres per al nostre conjunt de dades.

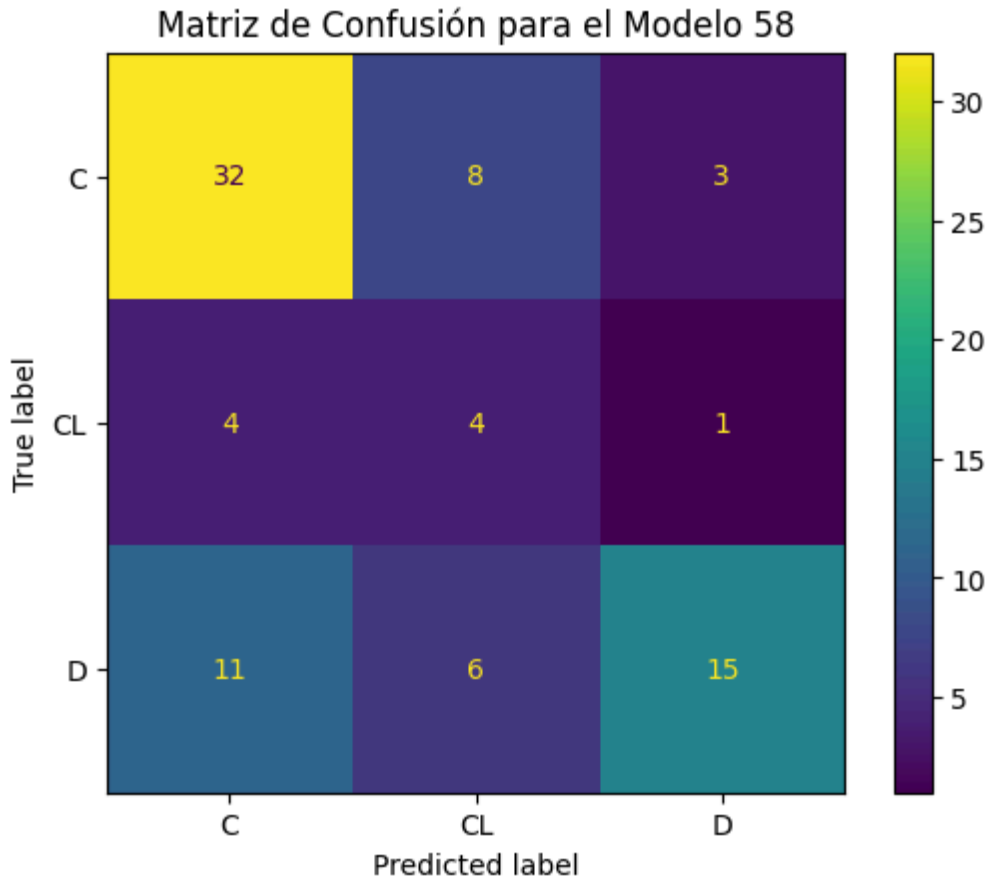
Selecció de Hiperparàmetres

- **n_neighbors (6-20):** Hem escollit explorar valors de *'n_neighbors'* entre 6 i 20. Aquest rang es justifica pel fet que tenim un conjunt de dades relativament petit. Amb pocs veïns (menys de 6), el model podria patir d'un sobreajustament, mentre que amb un nombre excessiu de veïns (més de 20), es corre el risc de sotaajustament. Aquest rang ens proporciona una bona balance entre capturar les subtileses locals de les dades i mantenir una bona capacitat de generalització.
- **weights:** Hem optat per utilitzar *'uniform'* com a valor per a aquest hiperparàmetre. Això significa que tots els veïns contribueixen igualment a la predicció. Tot i que la ponderació per distància ('distance') pot ser beneficiosa en alguns contextos, la simplicitat i transparència del pes uniforme són avantatjosos per a la interpretació i explicació dels resultats en un context clínic.
- **metric (['euclidean', 'manhattan', 'minkowski']):** La inclusió d'aquestes tres mètriques de distància ens permet explorar diferents maneres de mesurar la similitud entre els casos. La distància euclidiana és una elecció natural per a dades quantitatives, mentre que la distància de Manhattan pot ser més adequada per a dades amb distribucions atípiques. La distància de Minkowski ofereix una generalització d'aquestes dues, proporcionant una perspectiva més àmplia de la similitud.
- **algorithm (['ball_tree', 'kd_tree', 'brute']):** La selecció d'aquests tres algorismes ens permet avaluar quin és el més eficient per al nostre conjunt de dades. Els algorismes 'ball_tree' i 'kd_tree' són generalment més ràpids per a conjunts de dades grans, mentre que l'algorisme 'brute' pot ser més ràpid per a conjunts de dades més petits.

4.2.3 Entrenament

id	algorithm	metric	n_neighbors	weights	F1_Score
58	kd_tree	manhattan	7	uniform	0.767233
16	brute	manhattan	7	uniform	0.767233
100	ball_tree	manhattan	7	uniform	0.767233

4.2.4 Anàlisi de resultats



El model número 58, basat en l'algorisme k-nearest neighbors amb 7 veïns i la distància de Manhattan, mostra una competència notable per predir la mort dels pacients (classe 'D') amb un F1 Score de 0.767233 indicant un rendiment equilibrat en la nostra base de dades de cirrosi hepàtica. Tanmateix, el model es confon significativament entre pacients censurats per trasplantament de fetge ('CL') i censurats per altres raons ('C'), suggerint que la semblança en les seves característiques clíniques és un repte que requereix més discriminació. Això posa de manifest que elegir Knn no es un bon model per aquesta tasca.

4.3 Segon model (Arbre de decisions)

4.3.1 Motivació

Després de considerar el model KNN per les seves qualitats intuïtives i la seva eficàcia amb dades numèriques, ara ens centrem en l'Arbre de Decisions com a segon model per a la nostra investigació. Les característiques están explicados en "1.1.2Arbre de desicions".

En recordatori, l'Arbre de Decisions és escollit per la seva capacitat d'oferir una gran transparència en la presa de decisions, la seva habilitat per tractar diferents tipus de dades, i la seva competència en modelar la complexitat inherent de les dades clíniques. Aquests factors, combinats amb la seva facilitat d'ús i comprensió, el converteixen en un candidat sòlid per ser el segon model a explorar en la nostra recerca per a un model predictiu efectiu per a pacients amb cirrosi hepàtica.

4.3.2 Discussió dels hiperparàmetres disponibles, i dels valors usats

En el desenvolupament d'un model d'Arbre de decisions per a la predicció de la supervivència de pacients amb cirrosi hepàtica, la selecció acurada dels hiperparàmetres és essencial per optimitzar el rendiment del model. Com en el cas anterior amb el model KNN, hem utilitzat la tècnica de Grid Search per aquest propòsit. Aquesta elecció es basa en la mateixa raó: la falta de coneixement previ clar sobre quins serien els millors hiperparàmetres per a aquest conjunt de dades específic.

Selecció de Hiperparàmetres

- **max_depth: ([3, 5, 10, 15, None]):** La profunditat màxima de l'arbre (max_depth) és un factor clau per controlar el sobreajustament. Valors baixos (3, 5) poden prevenir un sobreajustament, però també poden portar a un sotaajustament, no capturant la complexitat dels patrons en les dades. Valors alts (10, 15) permeten que l'arbre capte complexitats més fines, però

incrementen el risc de sobreajustament. L'opció 'None' permet que l'arbre creixi fins que totes les fulles siguin pures o continguin menys que el mínim de mostres requerides per dividir-se, útil per explorar l'extrem de la complexitat.

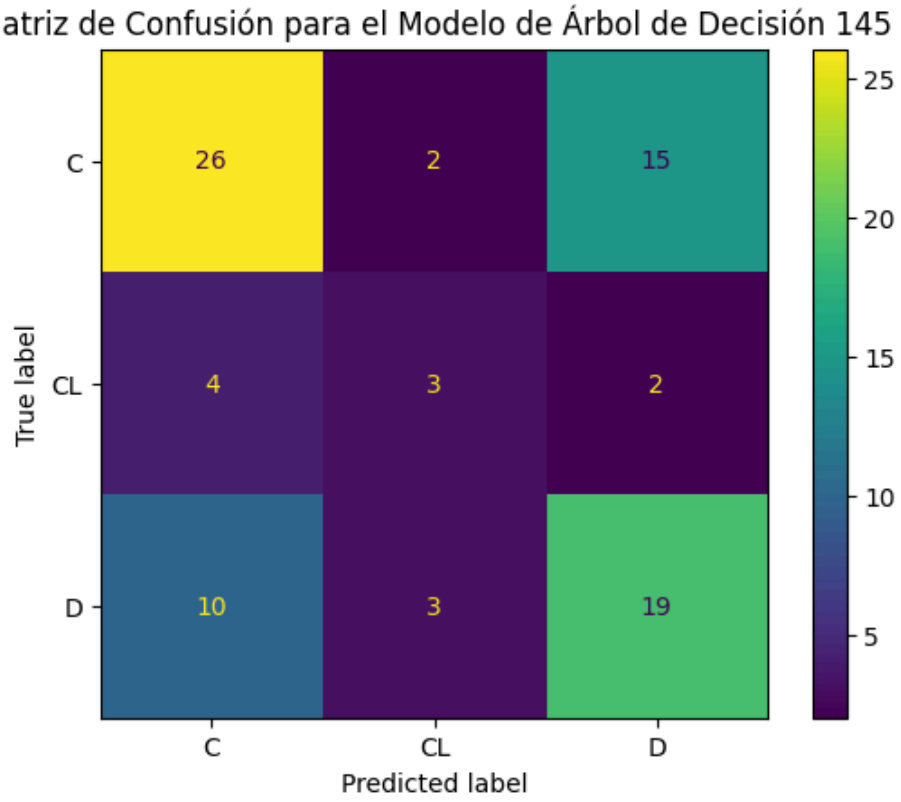
- **min_samples_split ([2, 5, 10]):** Aquest paràmetre determina el nombre mínim de mostres necessàries per dividir un node. Valors baixos (2, 5) faciliten que els nodes es dividixin, possiblement conduint a arbres més complexos i detallats. Un valor més alt (10) exigeix més mostres en un node per considerar la seva divisió, ajudant a simplificar l'arbre i a reduir el risc de sobreajustament.
- **min_samples_leaf ([1, 5, 10]):** Similar a min_samples_split, min_samples_leaf estableix el nombre mínim de mostres que ha de tenir una fulla. Valors més alts asseguren que l'arbre tingui fulles amb una quantitat més gran de dades, augmentant la fiabilitat de les prediccions i reduint el sobreajustament.
- **criterion (['gini', 'entropy']):** Aquests són criteris per mesurar la qualitat d'una divisió. 'Gini' és útil per a la seva rapidesa i eficiència, especialment en conjunts de dades grans. 'Entropy', basat en la teoria de la informació, pot ser més efectiu en identificar les millors divisions, però és lleugerament més lent. La comparació d'aquests dos criteris pot ajudar a determinar quin s'ajusta millor a les característiques específiques de les dades.
- **max_features (['sqrt', 'log2']):** Aquest paràmetre limita el nombre de característiques a considerar en cada divisió. 'sqrt' i 'log2' són estratègies comunes per reduir la variabilitat de l'arbre i augmentar la capacitat de generalització. Això pot ser particularment útil en conjunts de dades amb un gran nombre de característiques, per evitar l'ús excessiu de les mateixes característiques i per accelerar el procés de formació.

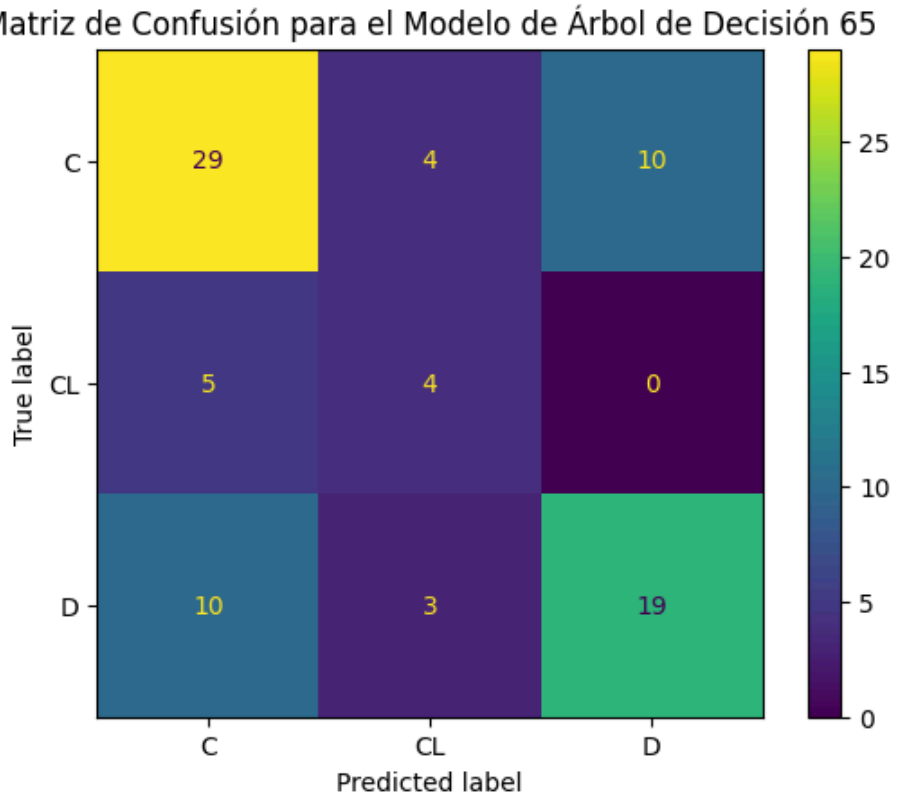
4.3.3 Entrenament

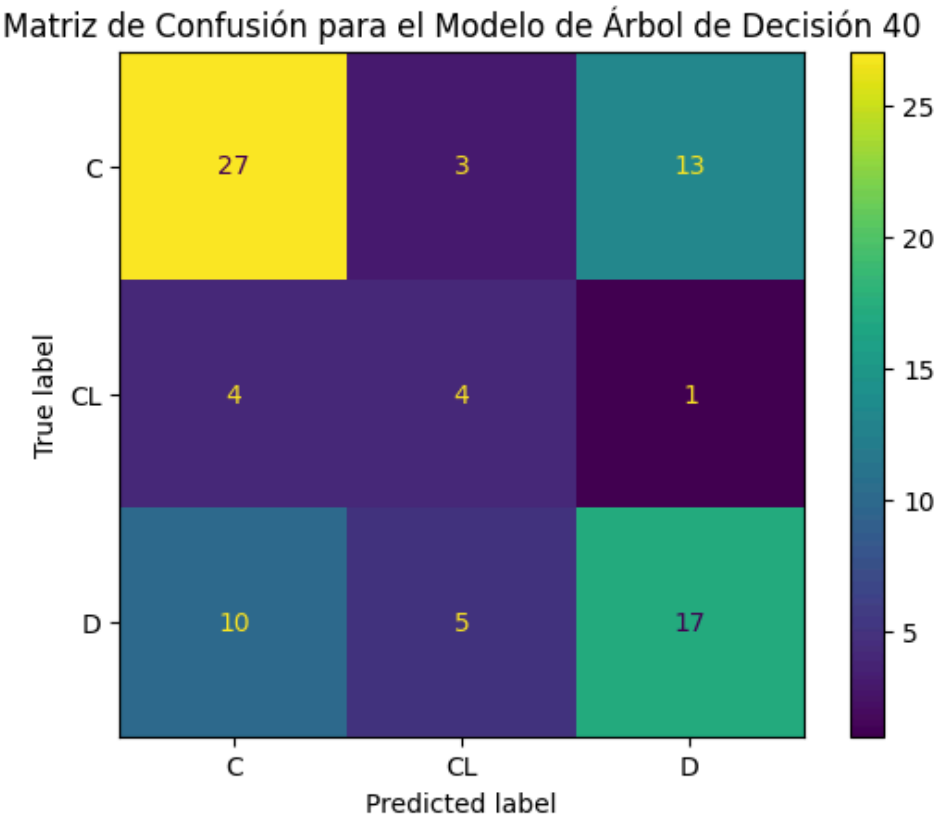
En aquest m'he donat compte que l'arbre de decisions de sklearn no accepta variables categòriques, així que per entrenar el model no vam fer ús de les dades X_train_balanced i X_test_imputed sinó que X_train_norm_est i X_test_norm_est.

id	criterion	max_depth	max_features	min_samples_leaf	min_samples_split	F1_Score
145	entropy	15.0	sqrt	1	2	0.813251
65	gini	15.0	log2	1	2	0.813201
40	gini	10.0	sqrt	5	5	0.804921

4.3.4 Anàlisi de resultats







En revisar les matrius de confusió dels models d'arbre de decisió, el model 65 sobresurt com el més efectiu, amb la millor taxa d'encerts en la classe 'C' (censurat) i un rendiment consistent en la predicció de la mort dels pacients (classe 'D'). Aquest model mostra una millor capacitat per equilibrar les prediccions entre les diferents classes, essent una opció preferible per a una aplicació clínica on és crucial distingir entre les diverses trajectòries de supervivència dels pacients amb cirrosi hepàtica.

4.2 Tercer model (SVM)

4.4.1 Motivació

Després d'analitzar els models KNN i Arbre de Decisions, orientem la nostra recerca cap al SVM (Support Vector Machine), un model potent i versàtil en el camp de l'aprenentatge automàtic. Les característiques están explicados en "1.1.3SVM".

Ateses les seves capacitats de modelatge sofisticades i la seva força en la classificació de dades complexes, SVM representa una opció sòlida per a la nostra tercera estratègia de modelatge.

4.4.2 Discussió dels hiperparàmetres disponibles, i dels valors usats

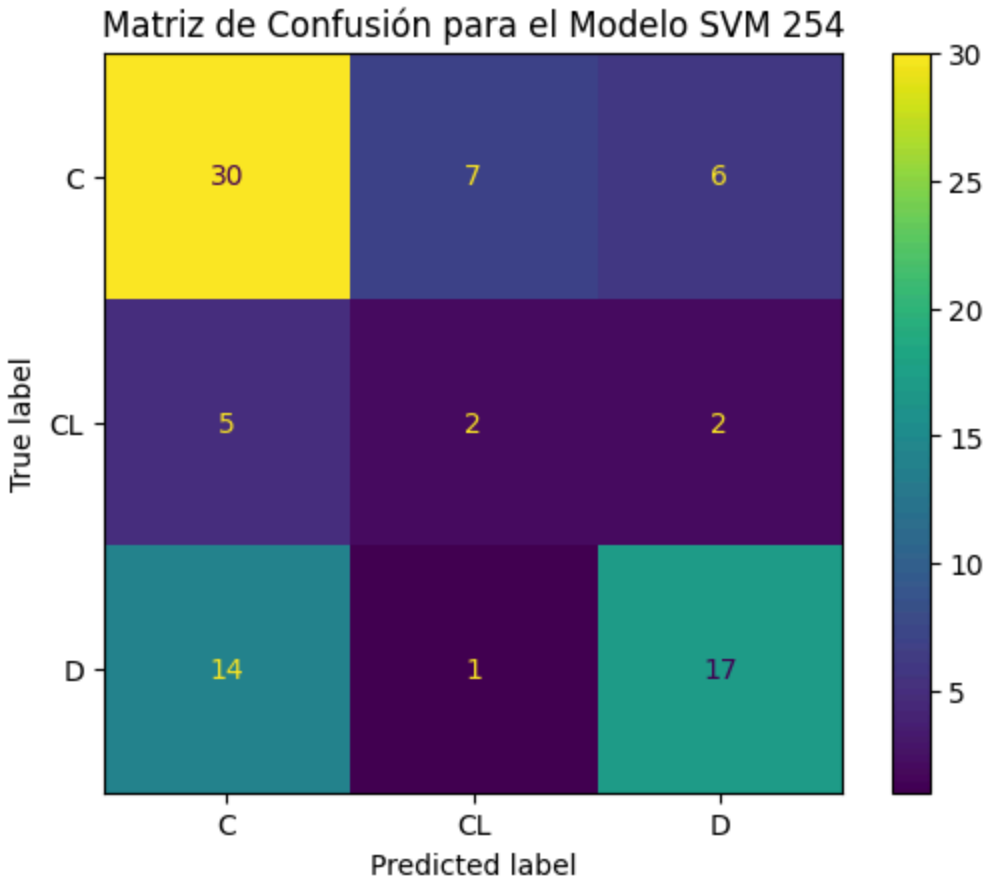
Selecció de Hiperparàmetres

- **C (1-20):** Hem limitat el rang de l'hiperparàmetre 'C' entre 1 i 25 per evitar un sobreajustament excessiu. Valors molt alts de 'C' poden fer que el model sigui massa sensible a les dades d'entrenament, perdent la capacitat de generalitzar bé a dades no vistes.
- **kernel (['linear', 'rbf', 'poly']):** La selecció del tipus de kernel 'linear', 'rbf' i 'poly' ens permet explorar diferents maneres de modelar les relacions entre les variables.
- **gamma (['scale', 0.1, 1]):** L'elecció de 'gamma' inclou 'scale' (que s'ajusta automàticament basat en el conjunt de dades), juntament amb valors fixos com 0.1 i 1. Aquesta gamma de valors ens permet investigar com el model reacciona a diferents graus d'influència dels exemples d'entrenament.

4.4.3 Entrenament

id	c	degree	gamma	kernel	F1_Score
254	10	3	scale	rbf	0.854270
245	10	1	scale	rbf	0.854270
263	10	4	scale	rbf	0.854270

4.4.4 Anàlisi de resultats



El model SVM 263 mostra una capacitat notable per a la identificació de pacients censurats (classe 'C'), amb una alta taxa d'encerts, però la seva eficàcia disminueix significativament en la predicció de la mort (classe 'D'), on es produeix una quantitat considerable de falsos negatius.

4.3 Comparació

Per avaluar quin model és el millor per predir la supervivència de pacients amb cirrosi hepàtica, es consideraran diverses mètriques de rendiment, com ara el F1 Score i l'anàlisi de les matrius de confusió.

Comencem analitzant les matrius de confusió:

1. **Model KNN (Model 58):**
    - Confusió significativa entre la classe D i les altres dues classes.
    - Molt bona classificació per la classe C.
2. **Model d'Arbre de Decisió (Models 145 i 65):**
    - Més equilibrat en la confusió entre classes en comparació amb el KNN.
    - Alta confusió entre la classe C i D, una situació no ideal quan es tracta de predir la supervivència dels pacients.
3. **Model SVM (Model 263):**
    - Similar al model d'arbre de decisió, però amb una millora en la distinció entre les classes C i D.
    - La confusió entre les classes C i CL és comparable a la de l'arbre de decisió.

Ara, mirant els millors F1 Scores de cada model:

- **KNN:** 0.767233 (id=58)
- **Arbre de Decisió:** 0.813251 (id=145)
- **SVM:** 0.854270 (id=263)

En base a aquestes dades, el model SVM presenta el millor F1 Score, el que suggereix que, en general, té un millor rendiment per predir correctament les classes, balancejant la precisió i el recall.

No obstant això, la tria del millor model també ha de tenir en compte la importància clínica de cada tipus d'error de classificació. Per exemple, si predir incorrectament la mort d'un pacient (classe D) és més greu que confondre si la censura és per un trasplantament hepàtic (CL) o no (C), això hauria de ser un factor clau en la decisió. Però això, no està especificat en l'enunciat.

5. SELECCIÓ DE MODEL

El model SVM id = 263, emergeix com la millor opció per a la tasca de predir la supervivència de pacients amb cirrosi hepàtica, ja que no només presenta el millor F1 Score, indicant un equilibri superior entre precisió i sensibilitat, sinó que també ofereix una matriu de confusió amb una distribució d'errors més favorable. Malgrat que els models d'arbre de decisió poden ser més explicatius, la superioritat quantitativa del SVM en aquest context específic és crucial, donada la importància de la precisió en les prediccions mèdiques.

5.1 Descripció del model triat

El model SVM (Màquines de Vectors de Suport) seleccionat opera amb un paràmetre de regularització, que indica un compromís moderat entre l'error de classificació i la simplicitat del model. Un C més alt pot implicar un model més complex amb un marge més petit, mentre que un C més baix pot conduir a un marge més ampli i un model més simple. Aquest valor suggeriria que el model busca un equilibri entre adaptar-se als dades d'entrenament i mantenir la generalització per a dades no vistes.

El paràmetre degree=3 es refereix al grau del polinomi en el cas que s'utilitzi el kernel polinòmic; tot i que en aquest cas, el kernel seleccionat és 'rbf' (radial basis function), aquest paràmetre no s'aplica.

gamma=scale indica que el coeficient per al kernel 'rbf' s'ajusta automàticament basat en el nombre de característiques en les dades, optimitzant així la transformació que el kernel 'rbf' aplica en el procés de mapeig de dades a un espai de característiques de major dimensió on els punts de dades poden ser linealment separables.

El kernel 'rbf' és particularment poderós en problemes de classificació complexos, ja que pot manejar casos on la relació entre les classes no és lineal, projectant les dades a un espai de característiques de dimensions més altes on es poden definir hiperplans per a la classificació.

En resum, el model SVM triat és un model robust que pot tractar relacions no lineals entre les característiques i la variable objectiu, buscant un equilibri entre l'error de classificació i la simplicitat del model, tot ajustant-se a les particularitats del conjunt de dades de supervivència de pacients amb cirrosi hepàtica.

5.2 Anàlisi de les limitacions i capacitats del model.

El model SVM triat, tot i ser potent en la classificació de les dades complexes, presenta algunes limitacions que cal considerar en un context clínic.

En primer lloc, la poca explicabilitat del model SVM és una preocupació significativa en àmbits on les decisions han de ser transparents i fàcilment interpretables, com és el cas de la medicina. La naturalesa matemàtica de com el model SVM crea fronteres de decisió no lineals fa difícil per als clínics interpretar la raó exacta darrere d'una predicció donada, el que pot ser un desavantatge en comparació amb models més interpretables com els arbres de decisió.

A més, encara que el F1 Score del model SVM sigui el més alt entre els models avaluats, un F1 Score de aproximadament 0.85 pot ser considerat insuficient en aplicacions de la vida real on les decisions poden tenir conseqüències greus. En el context de la prediccó de la supervivència dels pacients, un model ideal tindria un F1 Score molt a prop de 1, per assegurar que les decisions clíniques basades en aquestes prediccions siguin tan precises com sigui possible.

Pel que fa a les capacitats, els models SVM són coneguts per la seva alta eficàcia en espais de característiques grans i quan hi ha una clara marge de separació entre les classes. Això els fa adequats per a problemes de classificació com la predicció de la supervivència dels pacients, on les relacions entre les característiques poden ser altament no lineals i complexes. El kernel 'rbf' permet al SVM gestionar aquesta complexitat sense la necessitat de transformacions explícites de les característiques, cosa que pot ser molt útil quan s'aborden conjunts de dades mèdiques amb moltes dimensions.

6. MODEL CARD

1. Informació del Model:

- **Nom:** Model SVM per a Predicció de Supervivència en Pacients amb Cirrosi Hepàtica
- **Tipus:** Màquina de Vectors de Suport (SVM) amb Kernel Radial Basis Function (RBF)
- **Versió:** 1.0
- **Detalls:** Aquest model SVM està especialitzat en l'analítica predictiva de dades clíniques per a pacients amb cirrosi hepàtica, enfocat en la predicció de l'estatus de supervivència dels pacients: censurats (C), censurats a causa de trasplantament hepàtic (CL), o defunció (D).
-

2. Anàlisi i Preprocessament de Dades:

- **Preparació Pre-partició:** Inclou anàlisi univariant de variables categòriques, binàries i numèriques; estudi de balanceig de classes; detecció i tractament de valors faltants i outliers; i enginyeria de característiques amb construcció i selecció pertinent.
- **Gestió Post-partició:** S'ha realitzat la partició de dades, determinat la millor tècnica d'imputació per valors faltants, i s'ha aplicat tractament de balanceig amb una anàlisi profunda dels mètodes disponibles i la seva elecció final.
- **Codificació:** Transformació de variables categòriques per a la seva correcta interpretació pel model.
- **Anàlisi Post-processament:** S'ha realitzat una revisió detallada dels dades post-preprocessament per assegurar la qualitat i la consistència.
- **Preparació de Variables:** Normalització i escalat, anàlisi de correlacions, avaluació de variables categòriques respecte a la variable objectiu, eliminació de variables redundants o sorolloses, i estudi de la dimensionalitat amb PCA amb justificació de la seva no aplicació.

3. Desenvolupament del Model:

- **Dades Utilitzades:** La base de dades “Cirrhosis Patient Survival Prediction” extreta de l'UCI Machine Learning Repository, accessible a UCI Repository.
- **Paràmetres:** C=10, degree = 4, gamma=scale, kernel RBF.

- **Preprocessament:** Normalització, balanceig de classes, tractament de valors faltants i outliers, selecció de característiques.
- **Desenvolupadors:** Equip multidisciplinari en anàlisi de dades i intel·ligència artificial aplicada a la medicina.
-

4. Avaluació del Rendiment:

- **Mètriques de Rendiment:** Precisió, Recall, F1-Score. El model ha aconseguit un F1-Score de 0.854270.
- **Anàlisi de Rendiment:** Presentat en forma de matrius de confusió, indicant un millor rendiment en la classificació de la categoria de mort (D).

5. Ús Intencionat i Limitacions:

- **Ús Intencionat:** El model està destinat a ser utilitzat per professionals de la salut per a l'assistència en la presa de decisions clíniques.
- **Limitacions:** Menor explicabilitat per a la interpretació clínica directa, i un F1-Score que podria requerir millora per a una aplicació en la vida real sense supervisió experta.

6. Ètica i Conformitat:

- **Consideracions Ètiques:** Els dades utilitzats van ser recopilats amb consentiment informat i utilitzats de manera anònima.
- **Conformitat:** Model desenvolupat seguint els estàndards de privacitat i ètica, incloent GDPR.

7. Requeriments i Dependències:

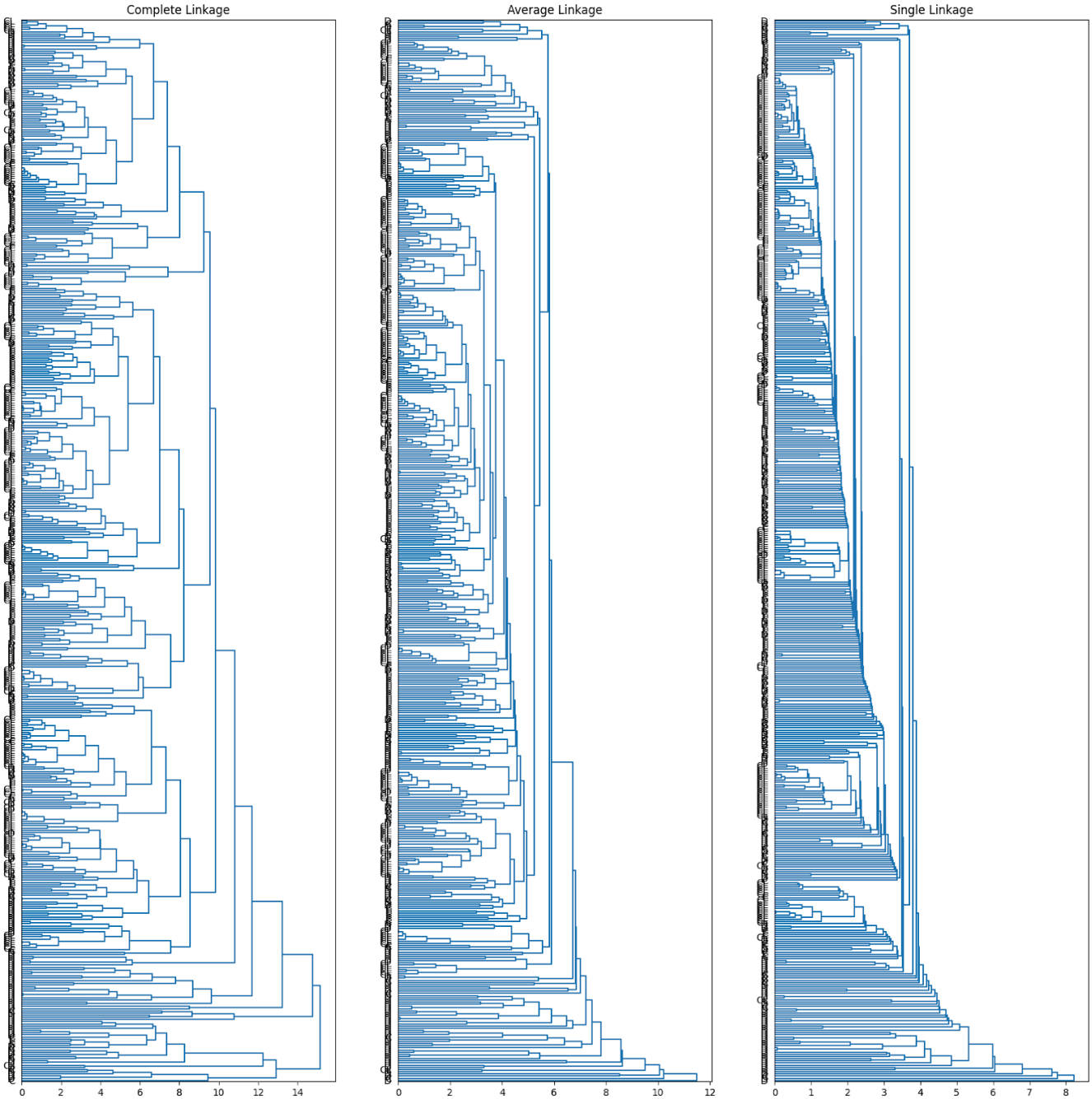
- **Requeriments Tècnics:** Python 3.6 o superior, llibreries de scikit-learn per a l'entrenament i avaluació del model.
- **Dependències:** Hardware amb capacitat de càlcul suficient per a l'entrenament i la inferència del model.

7. BONUS 2 (clusters)

7.1 Predir la supervivència dels pacients, tractant la variable 'Status'

Reducció de la Dimensionalitat mitjançant PCA

Inicialment, he aplicat l'anàlisi de components principals (PCA) per reduir la dimensionalitat de les dades. Això s'ha realitzat sobre X_train_norm_est, que presumiblement és un conjunt de dades normalitzades. L'objectiu d'aquest pas és simplificar l'estructura de les dades preservant al màxim la variància i la informació important. El resultat és un nou conjunt de dades transformades df2_plot, que s'escala i s'associa amb la variable Status de y_train_balanced_df. L'objectiu de això és poder realitzar un cluster jeràrquic.

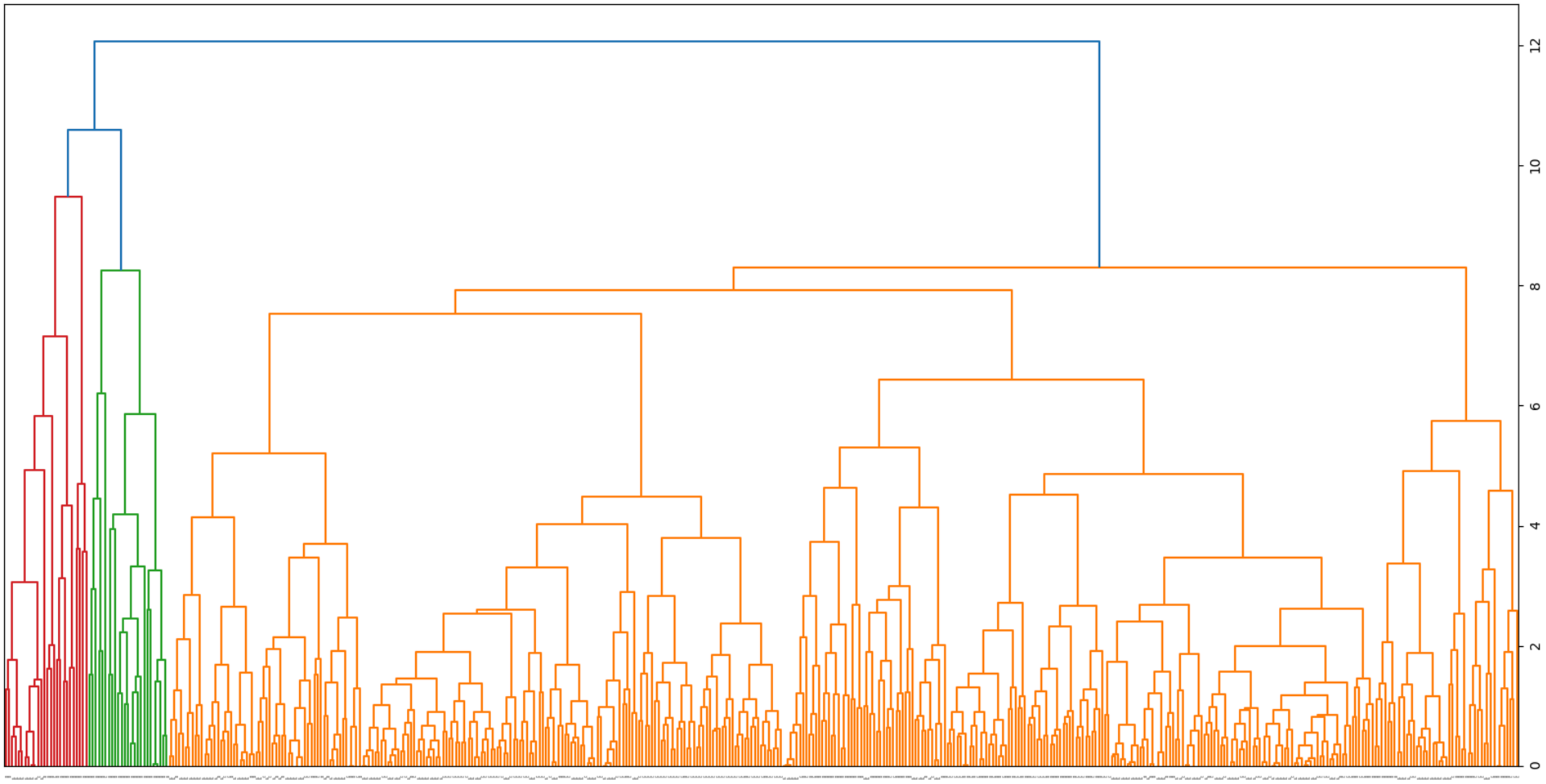

Complete Linkage
0
2
4
6
8
10
12
14
Average Linkage
0
2
4
6
8
10
12
Single Linkage
0
1
2
3
4
5
6
7
8

Dendrograms amb Diferents Mètodes d'Enllaç

Es pot observar com cada mètode agrupa les dades de manera diferent:

- **Enllaç Complet (Complete Linkage):** El dendrogram mostra clusters més definits i separats, indicant que aquest mètode pot ser més apropiat per identificar grups distints dins de les dades. És especialment útil quan les distàncies màximes entre els punts són rellevants per a la formació dels clusters.
- **Enllaç Mitjà (Average Linkage):** Els clusters semblen ser menys compactes que en l'enllaç complet, però més cohesius que en l'enllaç simple. Aquest mètode pren en compte la distància mitjana entre els punts de cada cluster, proporcionant un equilibri entre els mètodes complet i simple.
- **Enllaç Simple (Single Linkage):** Els clusters format per aquest mètode tendeixen a ser més allargats i menys compactes. Això es deu a que l'enllaç simple només considera la distància mínima entre els punts de diferents clusters, la qual cosa pot resultar en la unió de clusters basada en punts únics més que en l'estructura general dels grups.

A partir d'aquesta comparativa, he conclòs que l'enllaç complet és el mètode més adequat per a les dades.


0
2
4
6
8
10
12

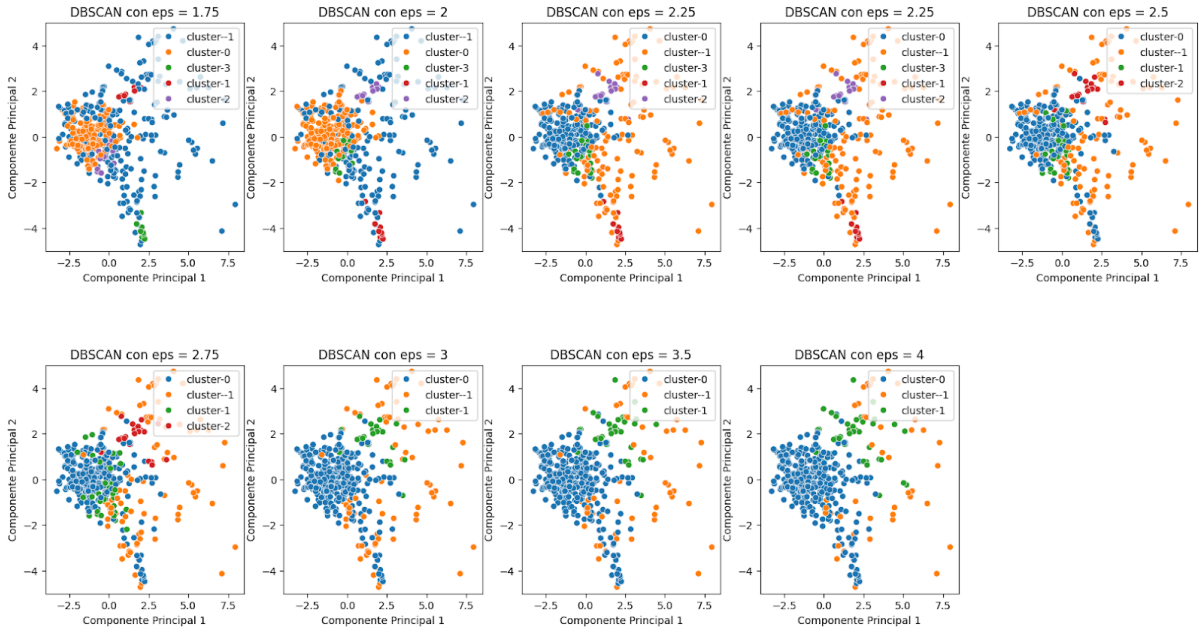
Anàlisi Detallada Utilitzant PCA i Enllaç Complet

Observem que les línies del dendrogram es coloregen diferent, probablement intentant reflectir els diferents estats de supervivència dels pacients. Tot i això, la sobreposició i proximitat dels colors podrien indicar que la separació entre els diferents estats de 'Status' no és tan clara com ens agradaria.

7.2 Clusters genèrics

En la secció anterior, 7.1, el meu objectiu era crear clusters jeràrquics que em permetessin identificar grups de pacients basant-me en la variable 'Status'. Malauradament, el resultat no va ser l'esperat; les dades no es van agrupar de manera clara i distintiva com havia anticipat. Això em va portar a la conclusió que un altre enfocament pot ser necessari.

tant, en aquesta nova secció, he decidit poder identificar diferents tipus de pacients independentment de la variable Status, aplicant un mètode de clustering diferent, DBSCAN .



Després d'aplicar DBSCAN i analitzar els resultats, he vist que, tot i els meus esforços, no he pogut extreure una classificació clara de les dades.