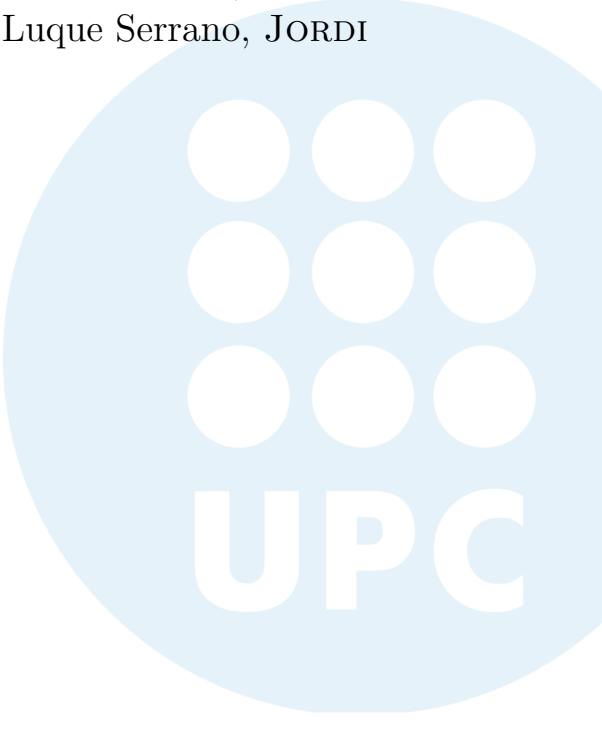
Facultat d'Informàtica de Barcelona

GIA UPC
Introducció a l'Aprenentatge Automàtic

Predicció d'estat de pacients amb cirrosi

Alumnes :
Granja i Bayot, Jordi

Tutors :
García Gasulla, Dario
Luque Serrano, Jordi



15 de març de 2024

Índex

1 Anàlisis i preprocessat de dades

1.1 Anàlisi estadstic de les variables de manera independent.

En aquesta base de dades trobem 20 variables inicialment. On, tot i que en un punt de partida estàn mal codificades, nou d'elles són categòriques i onze, numèriques segons la metadata. La variable categòrica ID però, la descartarem en primera instància i no passarà a l'anàlisi pel fet que és irrelevant identificar les instàncies i no té cap poder predictiu. Això clar, començarem amb l'anàlisi estadístic per variables categòriques i numèriques i finalment passarem a les distribucions

1.1.1 Anàlisi estadístic variables numèriques

	count	mean	std	min	25%	50%	75%	max
N_Days	334.0	1941.69	1129.26	41.00	1092.75	1743.50	2632.25	4795.00
Age	334.0	18455.76	3770.62	9598.00	15632.50	18628.00	21127.50	28650.00
Bilirubin	334.0	3.21	4.42	0.30	0.80	1.35	3.30	28.00
Cholesterol	235.0	373.78	237.39	127.00	251.50	315.00	402.00	1775.00
Albumin	334.0	3.50	0.42	1.96	3.26	3.54	3.77	4.64
Copper	256.0	97.61	89.36	4.00	39.75	73.00	123.25	588.00
Alk_Phos	258.0	2038.70	2214.00	289.00	944.00	1283.00	1949.75	13862.40
SGOT	258.0	122.75	57.02	26.35	80.95	116.25	151.90	457.25
Tryglicerides	234.0	121.77	56.41	33.00	84.00	108.00	151.00	382.00
Platelets	325.0	257.84	98.21	62.00	190.00	248.00	319.00	721.00
Prothrombin	334.0	10.73	1.04	9.00	10.00	10.60	11.10	18.00

Taula 1: Estadístic variables numèriques

Les observacions de la taula extretes a continuació com les de tot el treball, intenten generalitzar el caràcter de les diferents variables tenint en compte:

- **No tinc coneixement en el domini.**
- **Hi ha poques mostres, no puc extreure comportaments robustos.**
- **Desconec el grau del sampling bias del experiment.**
- **Les dades científiques són orientatives, cercades a la xarxa. Idòniament, es tindria el suport d'un expert.**
- **Tota observació feta és orientativa per guiar el procés d'anàlisi**

Observacions:

- **N_Days**: Variable de baixa informativitat a primera vista, ja que representa l'interval entre l'ingrés i la mort, trasplantament o finalització de l'estudi. La seva interpretabilitat varia entre pacients, limitant-ne la utilitat.

- **Edat**: Expressada inicialment en dies, la mitjana d'edat dels pacients amb cirrosi en aquest experiment és de 50 anys, amb un rang ampli de 25 a 78 anys, indicant una concentració significant en intervals específics (51-57 anys). Això suggereix que, un cop diagnosticada la cirrosi, la probabilitat d'edat adulta pot ser considerable.
- **Bilirrubina**: El rang segur de bilirubina (0,2 - 1,2 mg/dL) contrasta amb la mitjana observada de 3,21, més del doble de la fita superior. Més del 50% dels pacients superen aquesta fita, amb un màxim de 28, indicant una probabilitat significativa que, en cas de cirrosi, la bilirubina superi els límits saludables.
- **Cholesterol**: La mitjana de colesterol és de 373,78, molt per sobre del límit saludable de 240 mg/dL. Més del 75% de la població excedeix aquest límit, amb un màxim de 1775. Podem considerar una probabilitat elevada de nivells elevats de colesterol en pacients amb cirrosi. Ens trobem amb una desviació estàndard elevada.
- **Albumin**: Tot i que la mitjana d'albúmina és de 3,5, per sobre de la fita superior de normalitat (3,2 gm/dL), la majoria dels pacients es troben per sobre d'aquesta mitjana. Això indica que, en cas de cirrosi, la reducció d'albúmina no és una certesa, però les probabilitats poden augmentar en comparació amb la població general.
- **Copper**: Amb una mitjana de coure en orina de 97, més del 50% dels pacients superen el límit segur (15-60 ug/dia). Malgrat un màxim de 588, no podem descartar la seva veracitat. Vist que el pacient té cirrosi, podem veure un increment dels valors, sobrepassant el rang saludable.
- **Alk_Phos**: Els valors elevats de fosfatasa alcalina indiquen una possible variabilitat en les unitats de mesura o valors excepcionalment alts. Tot i això, la desviació estàndard elevada confirma la dispersió del conjunt de dades.
- **SGOT**: Amb una mitjana de 122,75, més del doble del límit superior (8-44 U/L), i la majoria de la població situada fora d'aquest rang, la cirrosi pot incrementar les probabilitats d'augment de SGOT. El màxim notable no pot ser ignorat, tot i que requereix una investigació més aprofundida dins de l'abast només d'un expert.
- **Triglicèrids**: Malgrat una mitjana de 121,77, per sota del límit perillós de 150 mg/dL, més del 50% dels pacients no presenten un excés significatiu. Amb un màxim de 382, encara que destacat, no es pot descartar.
- **Platelets**: Amb una mitjana de 257, dins de l'interval segur (150-450 platelets per cúbic), i només certa població fora d'aquests valors, no veiem cap possible conclusió deliberada.
- **Prothrombin**: Amb una mitjana de 10,73, per sota de la fita inferior segura (11-13,5 s), i més del 50% dels pacients situats per sota d'aquest interval, la cirrosi pot incrementar les probabilitats de baixos nivells de prothrombina. Malgrat que tenim un màxim de 18, no es pot descartar per desconeixement de domini.

1.1.2 Anàlisi estadístic variables categòriques

Variable	Count	Unique	Top	Frequency
Status	334	3	C	185
Drug	259	3	Placebo	129
Sex	334	2	F	298
Ascites	259	3	N	238
Hepatomegaly	259	3	Y	134
Spiders	259	3	N	185
Edema	334	3	N	282
Stage	330.0	4.0	3.0	123.0

Taula 2: Estadístic variables categòriques

Observacions:

- **Codificació:** Segons la metadata, les variables Drug, Ascites, Hepayomegaly, i Edema només tenen dues categories i les considerarem com a tal, ja que des de l'estudi no semblen donar-li cap valor informatiu a l'absència de la dada. Ens podem esperar un error de codificació.
- **Status:** Amb tres categories, la més freqüent és la C (viu sense trasplantament), que representa el 55% de les ocurrències. Aquest desequilibri és notable i cal tenir-ho en compte.
- **Drug:** Variable binària que presenta una distribució aparentment equilibrada, el que és coherent amb la naturalesa mèdica de l'experiment.
- **Sex:** Amb dues categories, la major part del sexe femení és evident, creant un desequilibri en comparació amb la població masculina.
- **Ascites:** Variable binària amb una clara majoria en la categoria N (no presència).
- **Hepatomegaly:** Variable binària on la categoria Y (si) és la més freqüent, tot i que presenta una distribució relativament equilibrada.
- **Spiders:** Variable binària amb una distribució aparentment equilibrada, destacant una majoria en la categoria N (no).
- **Edema:** Variable binària amb una notable diferència, amb una major freqüència de la categoria N (no).
- **Stage:** Variable amb quatre categories, on la categoria 3 predomina. Més endavant s'explorarà l'equilibri en les distribucions per aquesta variable específica.

1.1.3 Anàlisi descriptiu variables numèriques

En aquest apartat observarem les distribucions i boxplots de les variables numèriques de la nostra base de dades. Intentarem extreure informació valuable que ens permetrà interpretar les nostres dades de forma acertada. S'ha utilitzat ja la distinció per variable objectiu en els histogrames per facilitar explicacions en un futur.

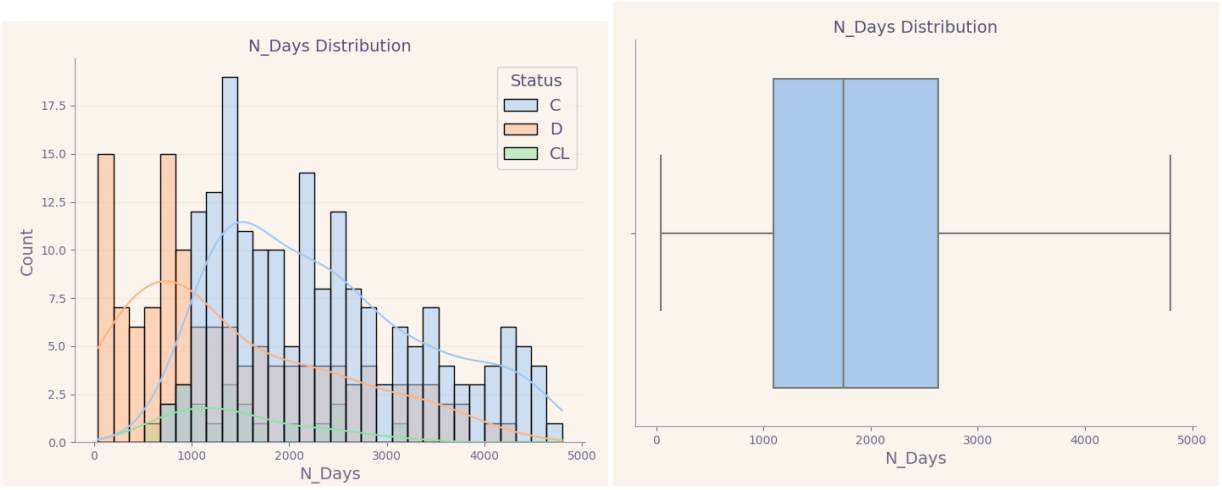


Figura 1: Histograma i boxplot variable N_Days

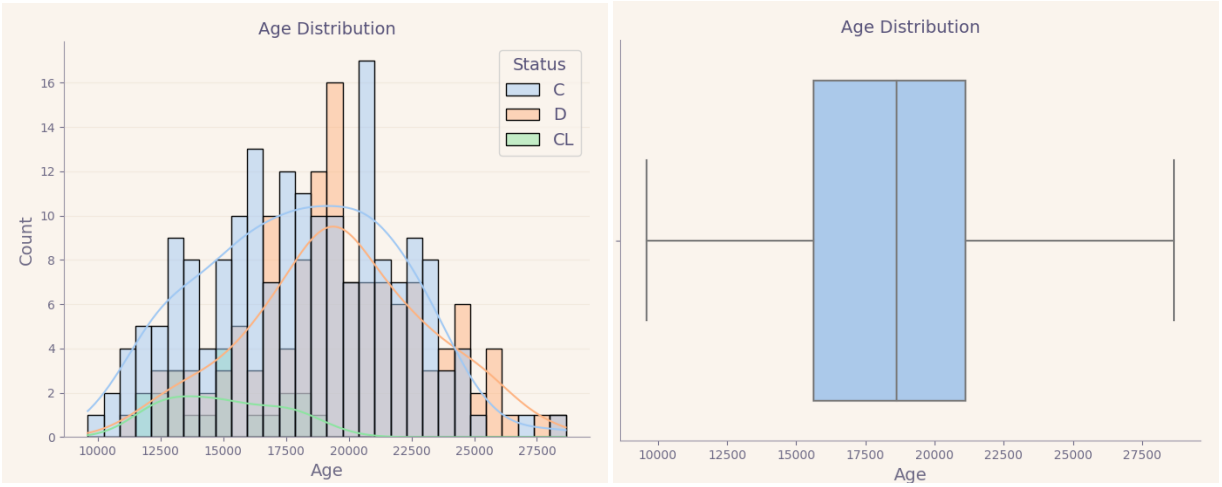


Figura 2: Histograma i boxplot variable Age

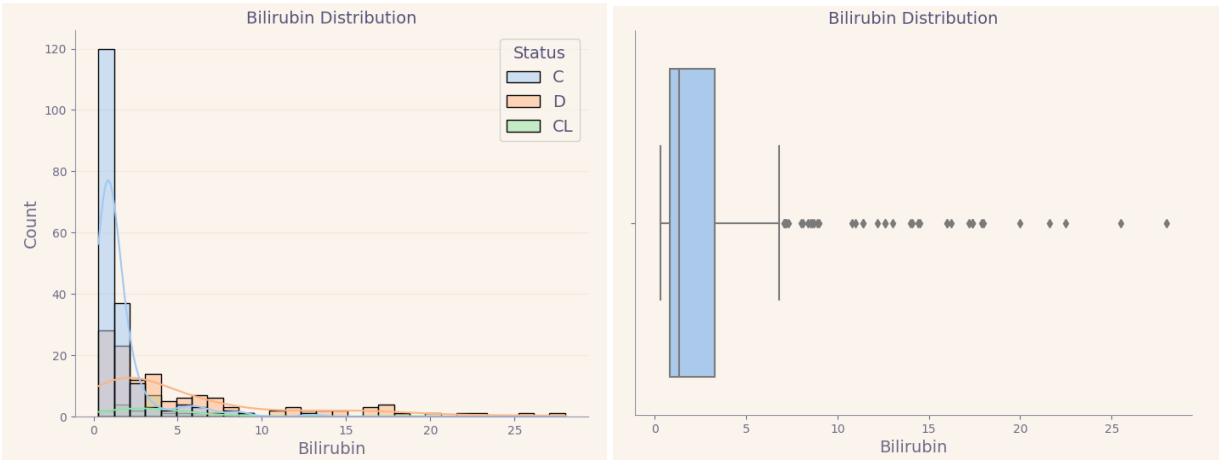


Figura 3: Histograma i boxplot variable Bilirubin

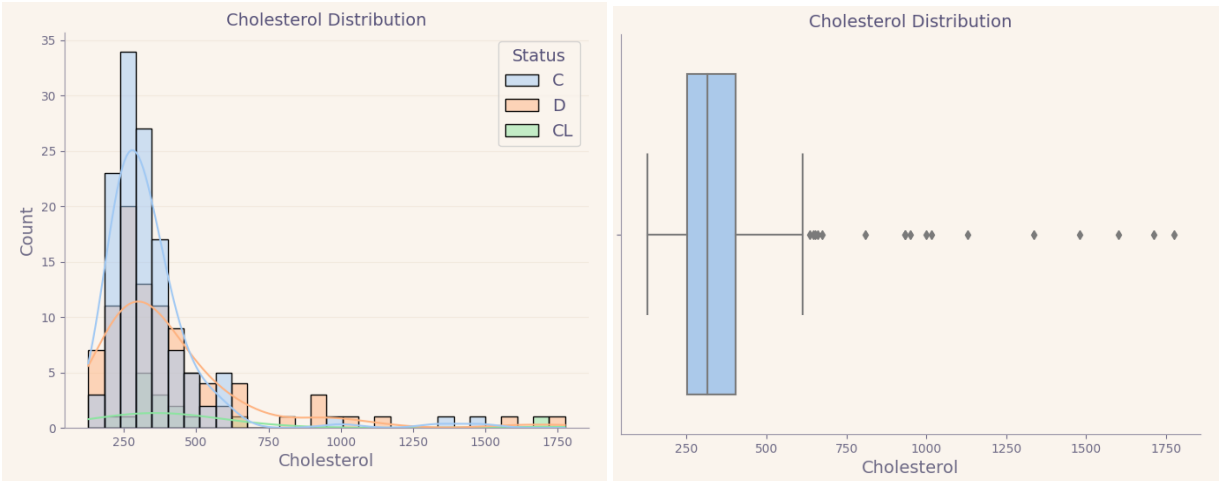


Figura 4: Histograma i boxplot variable Cholesterol

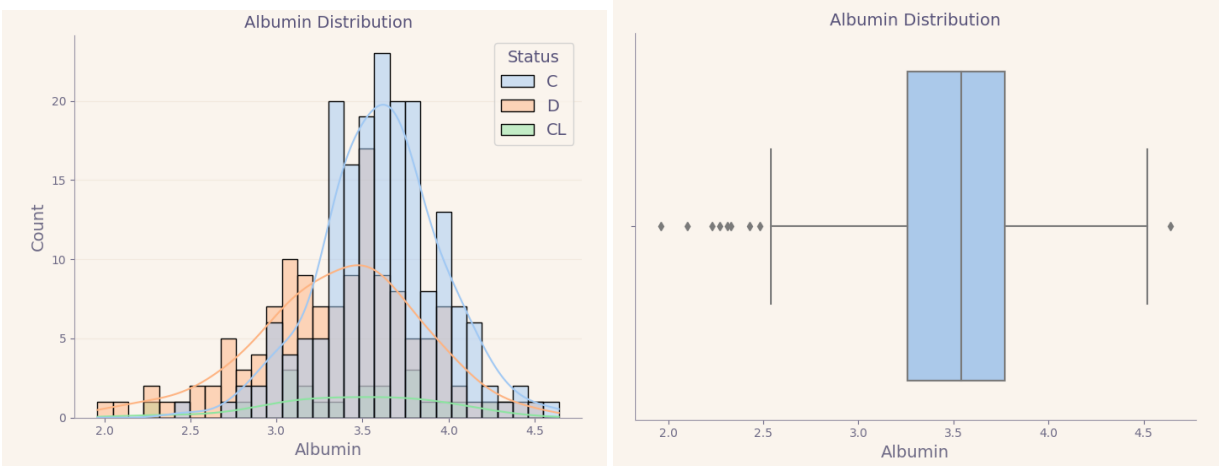


Figura 5: Histograma i boxplot variable Albumin

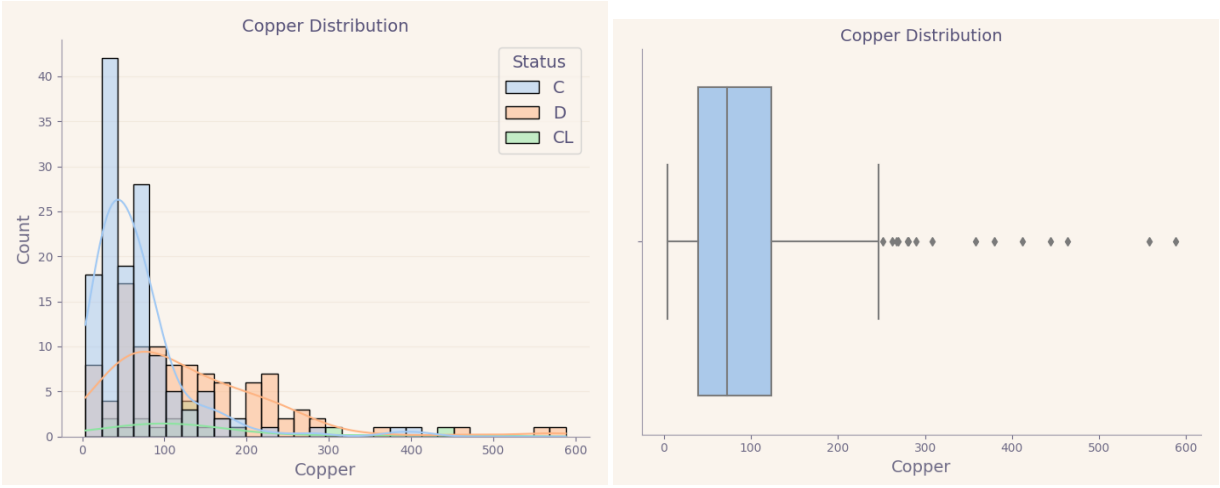


Figura 6: Histograma i boxplot variable Copper

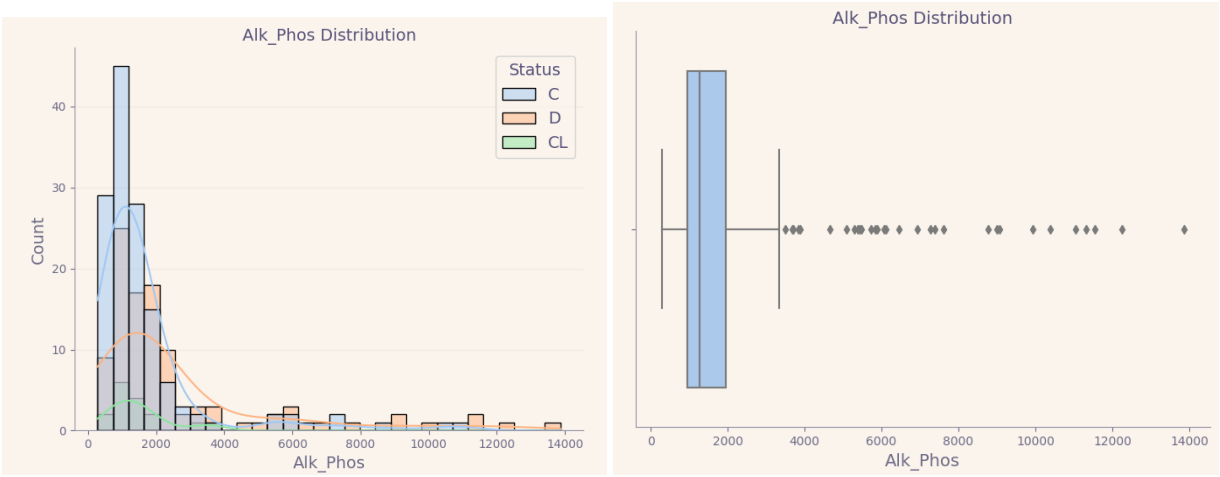


Figura 7: Histograma i boxplot variable Alk_Phos

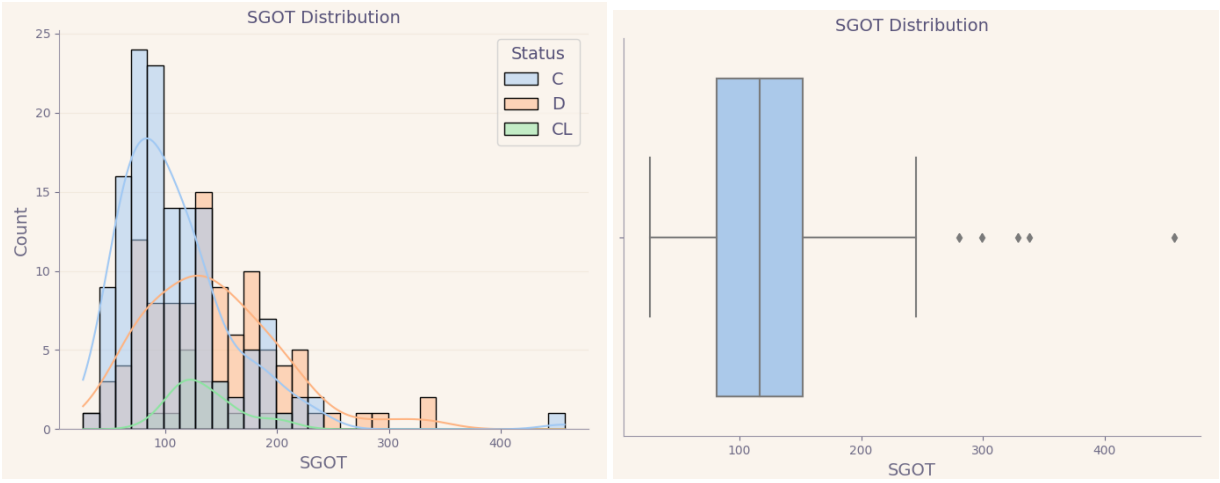


Figura 8: Histograma i boxplot variable SGOT

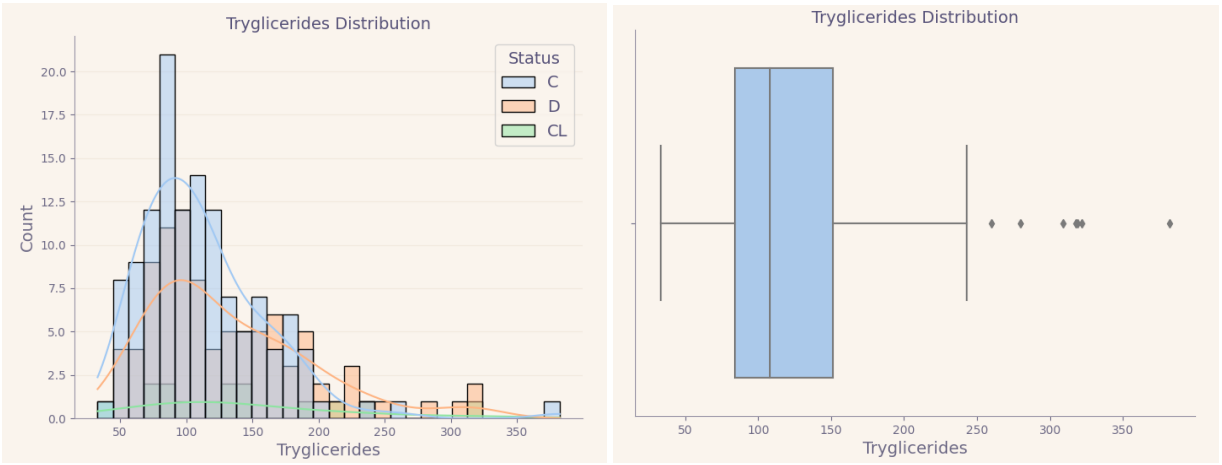


Figura 9: Histograma i boxplot variable Tryglicerides

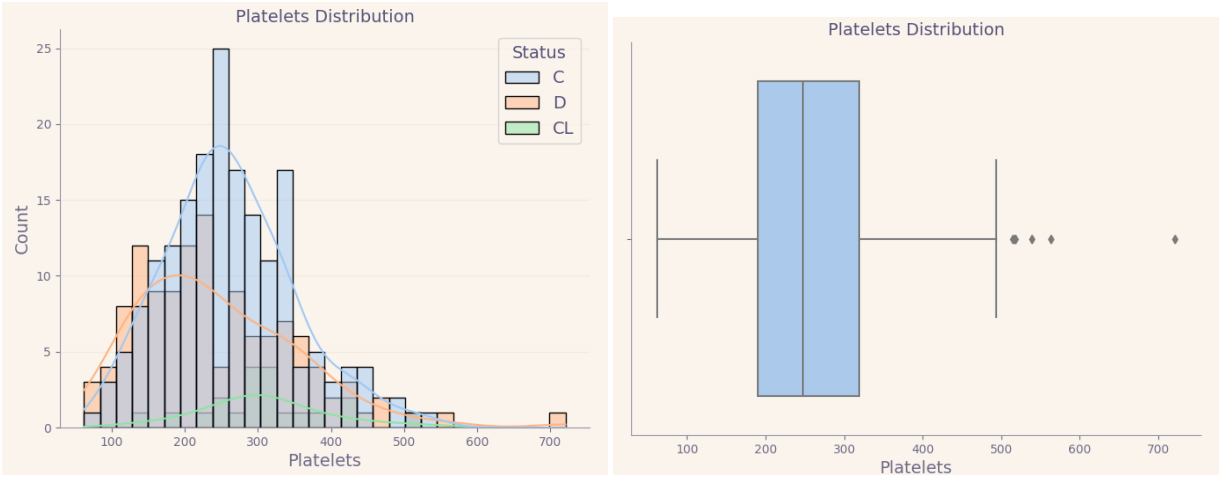


Figura 10: Histograma i boxplot variable Platelets

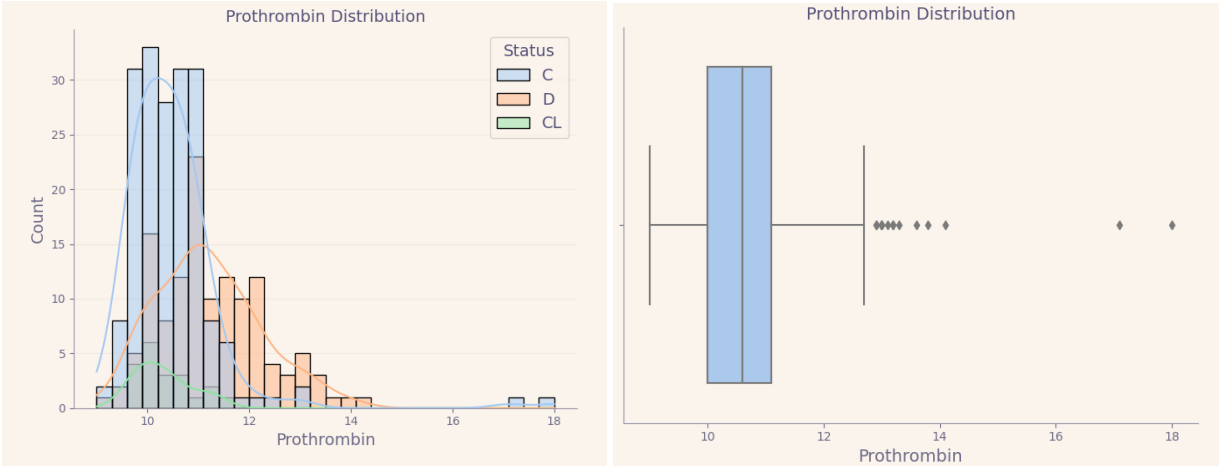


Figura 11: Histograma i boxplot variable Prothrombin

Observacions :

- **N_Days**: Distribució aproximable a una normal però lleugerament multimodal ja que podem veure diversos pics. No observem potencials outliers i podem observar que la distribució està lleugerament desviada concentrada a la part esquerra.
- **Edat**: Distribució de natura normal, sense outliers significatius, i centrada.
- **Bilirrubina**: Distribució de natura Power-Law. Trobem alguns potencials outliers molt més enllà del tercer quartil. Concentra la majoria de les instàncies en les primeres unitats com és comú en aquest tipus de distribucions.
- **Cholesterol**: Distribució de natura quasi-normal amb molta cua. Trobem bastants potencials outliers. Centrada a l'esquerra.
- **Albumin**: Distribució de natura normal amb cua als valors petits. Centrada a la dreta amb algun outlier a la part inicial.
- **Copper**: Distribució de natura Power-Law centrada a l'esquerra amb potencials outliers.
- **Alk_Phos**: Distribució de natura Power-Law centrada a l'esquerra amb potencials outliers.
- **SGOT**: Distribució de natura normal centrada a l'esquerra amb cua a la dreta. Ens trobem amb alguns potencials outliers.
- **Triglicerides**: Distribució de natura normal centrada a l'esquerra amb cua a la dreta. Ens trobem amb alguns potencials outliers.
- **Platelets**: Distribució de natura normal centrada a l'esquerra amb cua a la dreta. Ens trobem amb alguns potencials outliers.
- **Prothrombin**: Distribució de natura normal amb una gran concentració entre els valors 9 i 11. Centrada a l'esquerra amb cua a la dreta. Ens trobem amb dues instàncies molt distants de la població.

1.1.4 Anàlisi descriptiu variables categòriques

En aquest apartat veurem les freqüències per categoria de les variables categòriques. Les observacions que podem extreure'n, però, coincideixen amb les de l'apartat **1.1.2: Anàlisi estadístic variables categòriques**. S'ha eliminat prèviament les males codificacions comentades 'NaNN' per claredat:

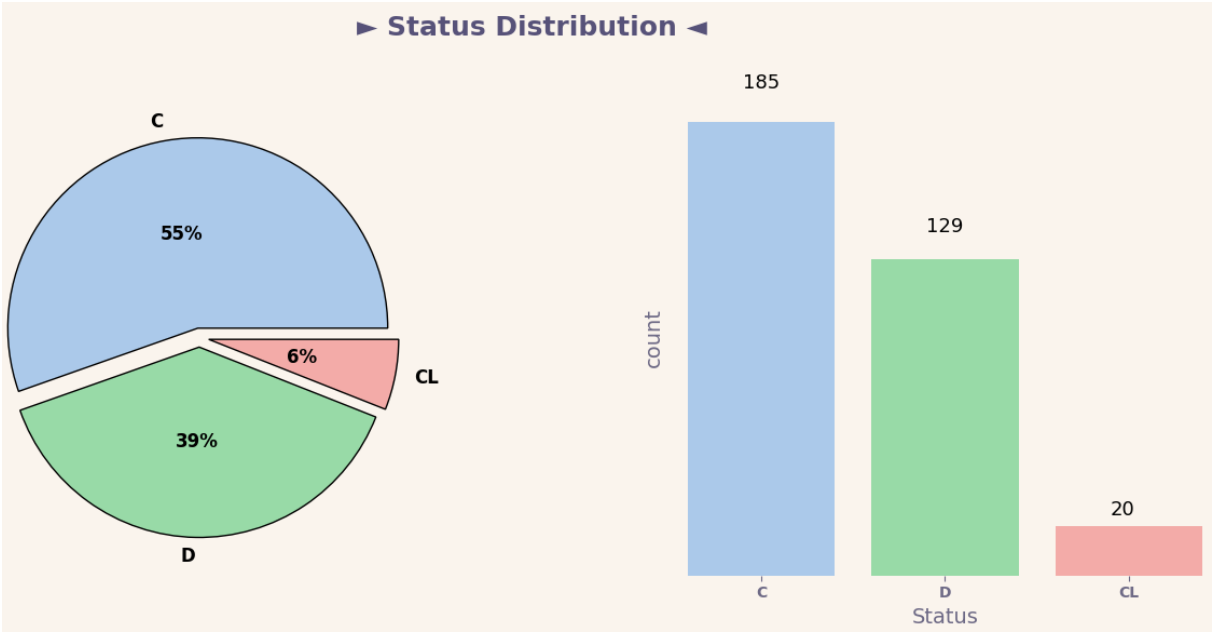


Figura 12: Countplot i piechart variable objectiu Status

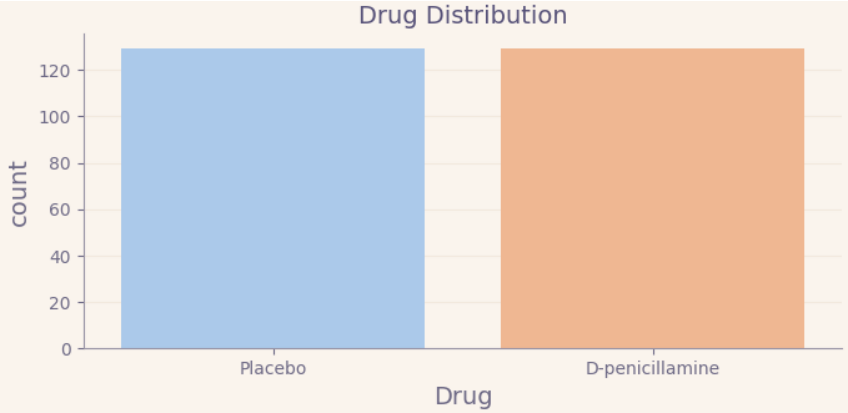


Figura 13: Countplot variable Drug

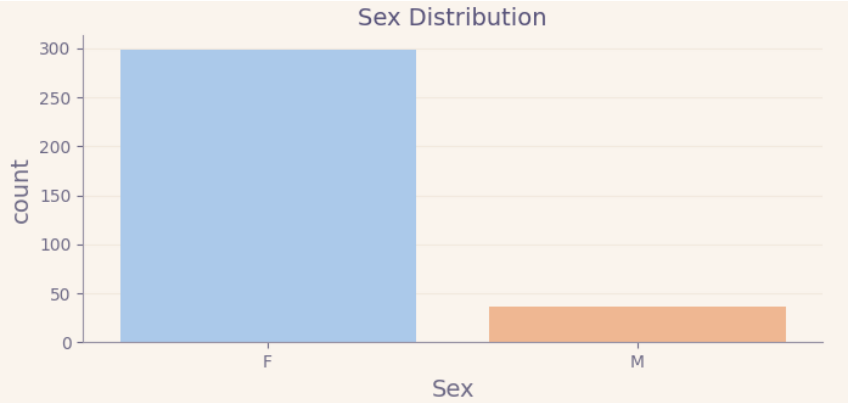


Figura 14: Countplot variable Sex

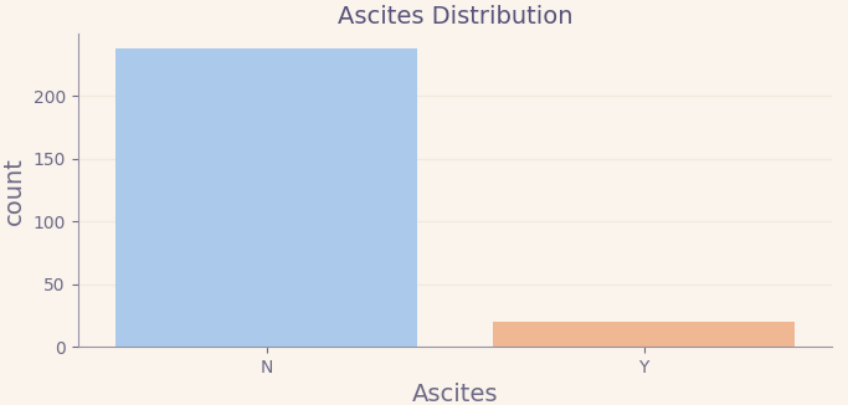


Figura 15: Countplot variable objectiu Ascites

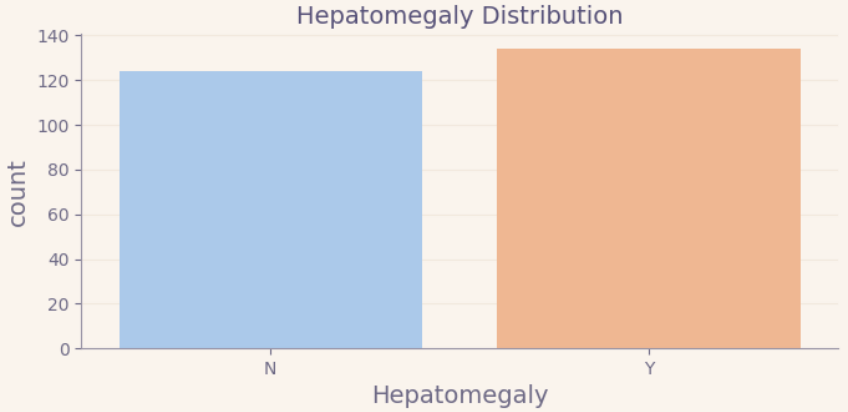


Figura 16: Countplot variable Hepatomegaly

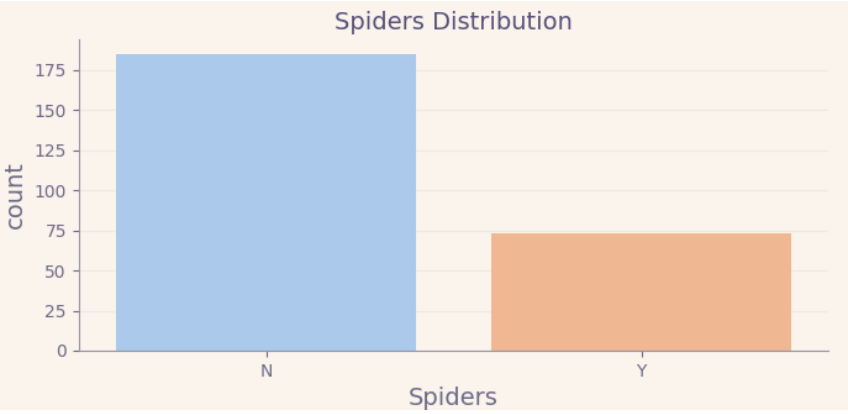


Figura 17: Countplot variable objectiu Spiders

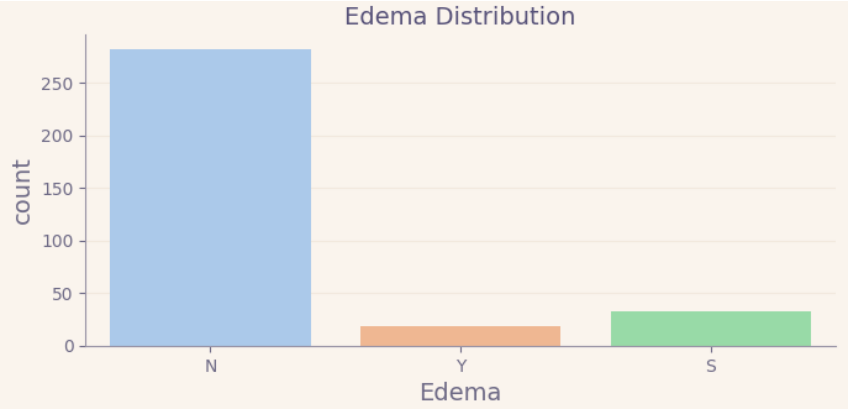


Figura 18: Countplot variable Edema

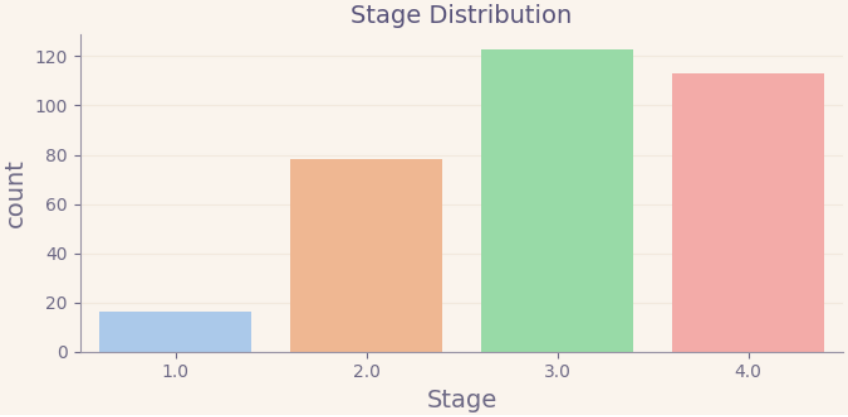


Figura 19: Countplot variable Stage

Hem pogut observar el comentat anteriorment, amb especial atenció però en el desequilibri de la classe objectiu.

1.2 Estudi de balanceig de classes

En aquest apartat valorarem el balanceig de la variable objectiu **Status** així com el seu tractament de cara al modelatge a posteriori.

Cal destacar que tot el procés d'EDA(Exploratory Data Analysis), tant en aquest apartat com els anteriors i els futurs, s'ha realitzat després del train/test split tot i que s'ha fet una ràpida visualització de les distribucions abans (procés present al notebook). Les raons per les quals he procedit d'aquesta manera són:

- **Modelatge més precís del món real:** En el món real només tindria accés a les dades d'entrenament.
- **Prevenir el data leakage:** Aquest procés implica explorar les relacions i patrons dins de les dades. Si ho fes en tot el conjunt de dades abans de dividir-lo, correria el risc d'extreure informació involuntàriament del conjunt de proves, la qual cosa provocarà una fuita de dades.

La freqüència per classe de la variable objectiu, ja vista, és de la forma següent:

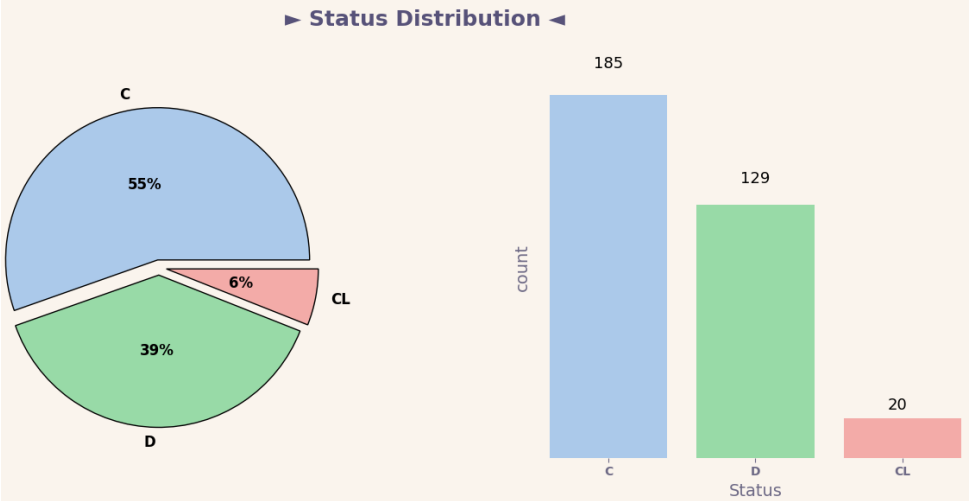


Figura 20: Countplot i piechart variable objectiu Status

Podem observar un desequilibri entre les tres classes que pot provocar un bias cap a les classes majoritàries (C i CL) en els diferents models, mala generalització, o fins i tot confusió en mètriques com l'accuracy.

Per a afrontar aquest problema tenim tres alternatives en el context del nostre problema: donar importància a la funció de cost (1), realitzar undersampling (2), i finalment realitzar oversampling (3). Parlarem dels avantatges i desavantatges de les diferents alternatives.

Mètode	Pros	Contres
1	No cal generar dades sintètiques ni reduir el nombre d'instàncies.	No podrem utilitzar-lo en models sense funció de cost (KNN), estem intervenint en el model directament.
2	No cal generar dades sintètiques.	Ens quedem amb tantes mostres com la classe minoritària (16), inadmissible.
3	Tindrem un model equilibrat, no esbiaixat, i amb tendència a generalitzar millor.	Estem generant quasi 200 instàncies sintètiques partint de 20 mostres de la classe minoritària. Més les de la segona classe majoritària.

Tot i que en el millor dels casos, amb una altra base de dades no buscaríem cap de les alternatives, s'ha de remarcar que estem treballant amb una base de dades amb molt poques instàncies i amb un desequilibri considerable. Ja que s'ha de prendre una decisió i veient que el mètode 1 i 2 tenen desavantatges majors o inconvenients, optarem pel tercer mètode en primera instància. Tot i aquesta primera decisió, a l'apartat **3: Definició de Models** fem un experiment per a decidir quin mètode utilitzarem, el qual dona per victoriós els 'weightings'.

Per a poder realitzar diagnosi del model a posteriori, cal caracteritzar les conseqüències que té aplicar oversampling, concretament el mètode SMOTE (tot i que en la creació de models valorarem mètodes com el bootstrapping), sobre la nostra base de dades:

- **Overfiting:** SMOTE pot causar overfiting, especialment en els casos on es generen moltes mostres a partir de poques.
- **Diversitat limitada**: En conjunts de dades petits, pot haver-hi una diversitat limitada en la classe minoritària, i SMOTE pot generar mostres sintètiques que estan massa a prop dels exemples existents, resultant en un augment menys eficaç.

Tot i això, més endavant en la creació de models, faré comparatives de mètriques entre el mètode 1,2,3 explicats anteriorment per reforçar la decisió d'aquest apartat.

1.3 Missings

En aquest apartat identificarem i proposarem la gestió de missings de la nostra base de dades. En primer lloc, mirarem el percentatge de missings per columna:

Variable	#	**Missing Values (%)**
Tryglicerides	100	29.94
Cholesterol	99	29.64
Copper	78	23.35
SGOT	76	22.75
Alk_Phos	76	22.75
Drug	75	22.46
Ascites	75	22.46
Hepatomegaly	75	22.46
Spiders	75	22.46
Platelets	9	2.69
Stage	4	1.20
Prothrombin	0	0.00
ID	0	0.00
Albumin	0	0.00
N_Days	0	0.00
Edema	0	0.00
Sex	0	0.00
Age	0	0.00
Status	0	0.00
Bilirubin	0	0.00

Com es pot apreciar, hi ha un gran percentatge de missings en algunes columnes de la nostra base de dades.

El primer que cal analitzar és la natura dels missings. Tenen aquests missings significat en alguna de les columnes? El metadata no dona explicació, així assumint que no en tenen. Per tant, els caracteritzarem com a 'random' i en considerarem la imputació. Són aquests missings localitzables? La metadata ens diu que s'ha annexat 112 pacients de fora l'experiment per a les mètriques bàsiques on, si ens fixem en aquests individus, veiem que tenen fins a 9 o 10 columnes buides. Aquestes xifres les podem veure a la figura següent, la qual mostra el recompte de missings de les diferents instàncies:

Figura 21: Countplot nombre de missings

Un cop caracteritzat el primer problema, hem de prendre la decisió d'eliminar (1) o no (2) aquestes instàncies tenint en compte el que això comporta:

Mètode	Pros	Contres	Implicacions
1	Reduïm considerablement el nombre de missings.	Perdem pràcticament un terç de les mostres en un dataset molt reduït, perdem capacitats predictives de tota mena.	Si es redueixen molt els missings podem utilitzar mètodes d'imputació relativament lleugers.
2	Conservem 112 pacients en un dataset de tan sols 312 mostres, pacients que si es tracten de manera adequada poden contribuir a la qualitat del model.	Tindrem pacients bastant sintètics que poden no representar la població general.	Haurem d'usar mètodes d'imputació relativament potents com pot ser KNNImputer o MICE(IterativeImputer) perquè tenim un gran percentatge de missings.

Vist això, procedirem amb el mètode 2. Pel fet que a primera vista no ens convé reduir un dataset tan petit en un terç.

Una primera preocupació que ve després dels apartats de balanceig i missings, és l'adopció de mètodes per crear dades sintètiques com pot ser SMOTE per sobre de dades majoritàriament sintètiques a causa de la gran quantitat de missings. Tot i això, procedirem amb les decisions preses, i intentarem reforçar els seus arguments amb resultats.

Cal mirar, doncs, com queden les distribucions pre i posttractament de missings tant per les variables categòriques i numèriques:

Pel que fa a les variables categòriques que mostraven missings:

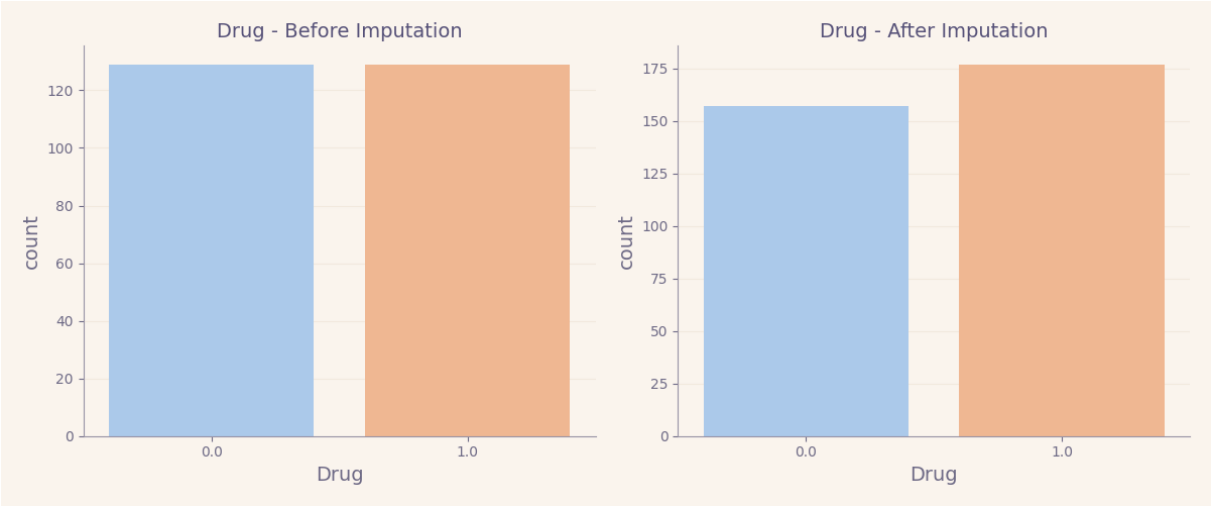


Figura 22: Countplot tractament missings variable Drug

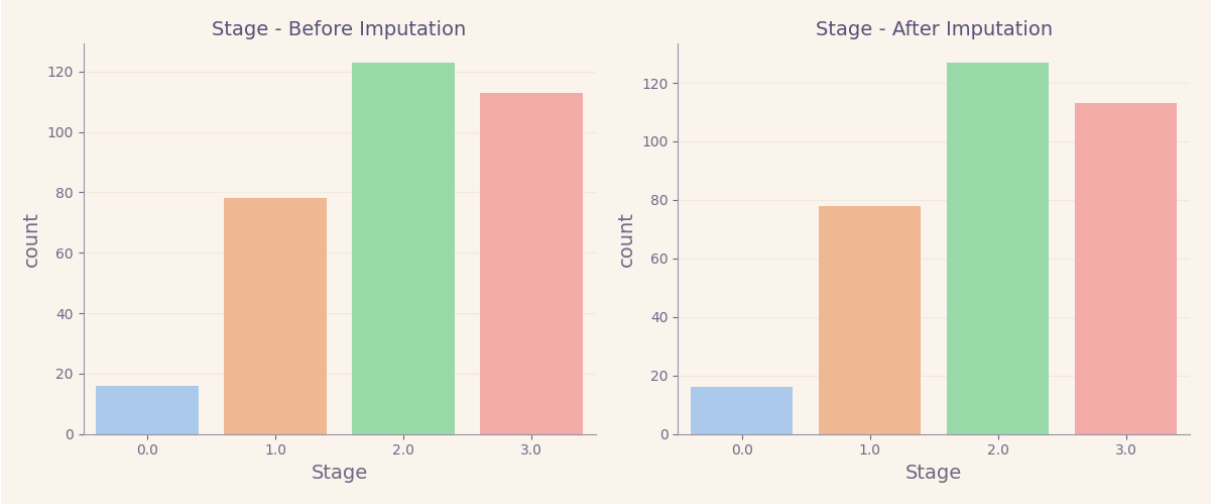


Figura 23: Countplot tractament missings variable Stage

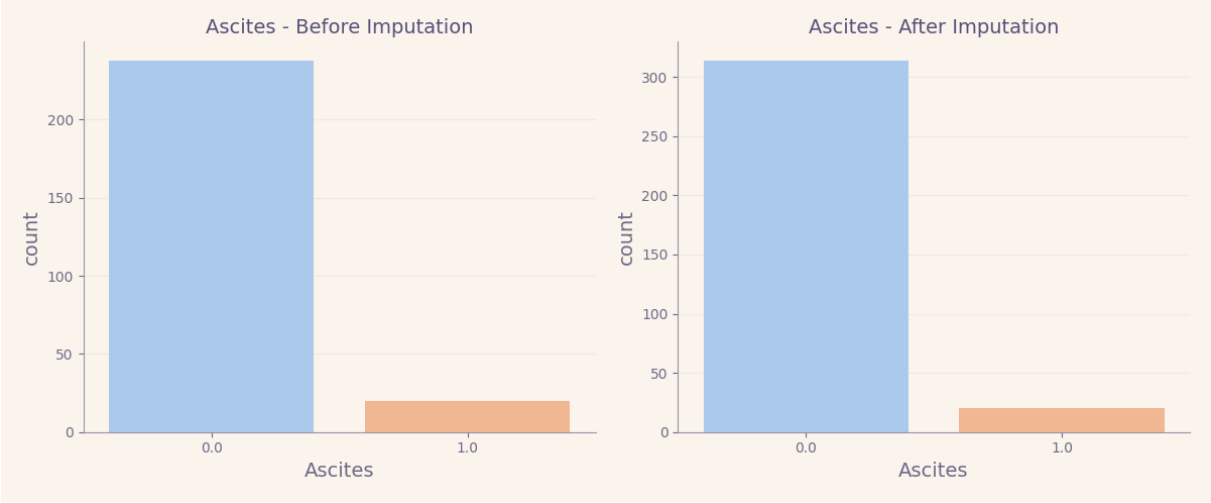


Figura 24: Countplot tractament missings variable Ascites

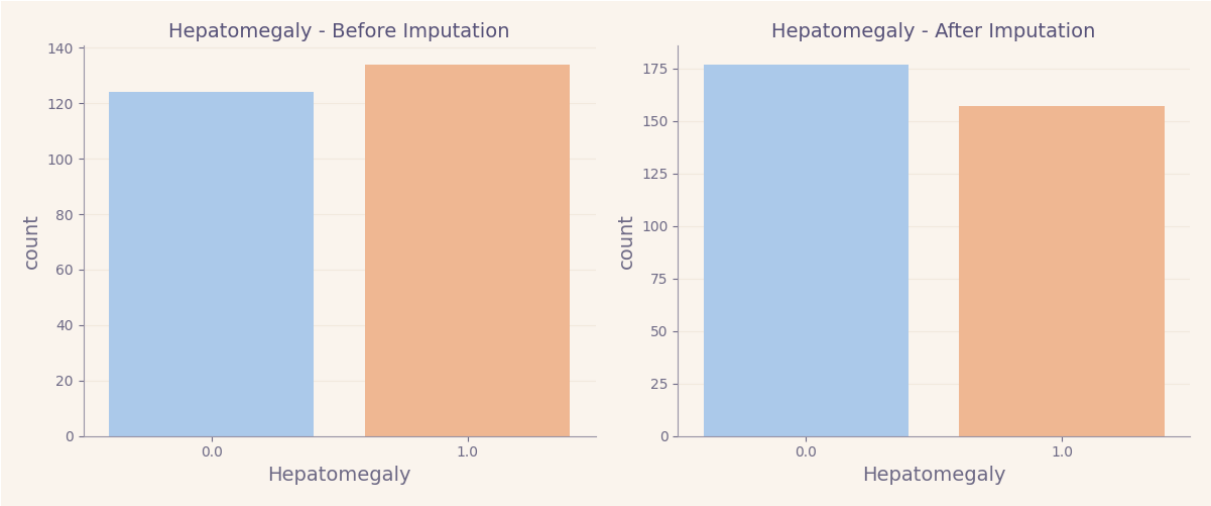


Figura 25: Countplot tractament missings variable Hepatomegaly

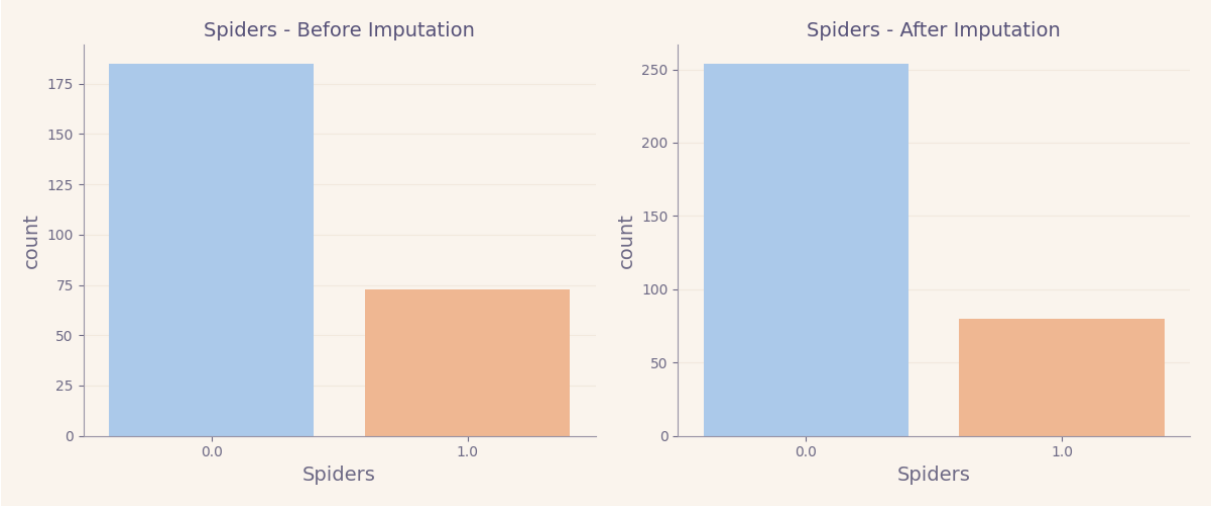


Figura 26: Countplot tractament missings variable Spiders

En quant a les variables numèriques:

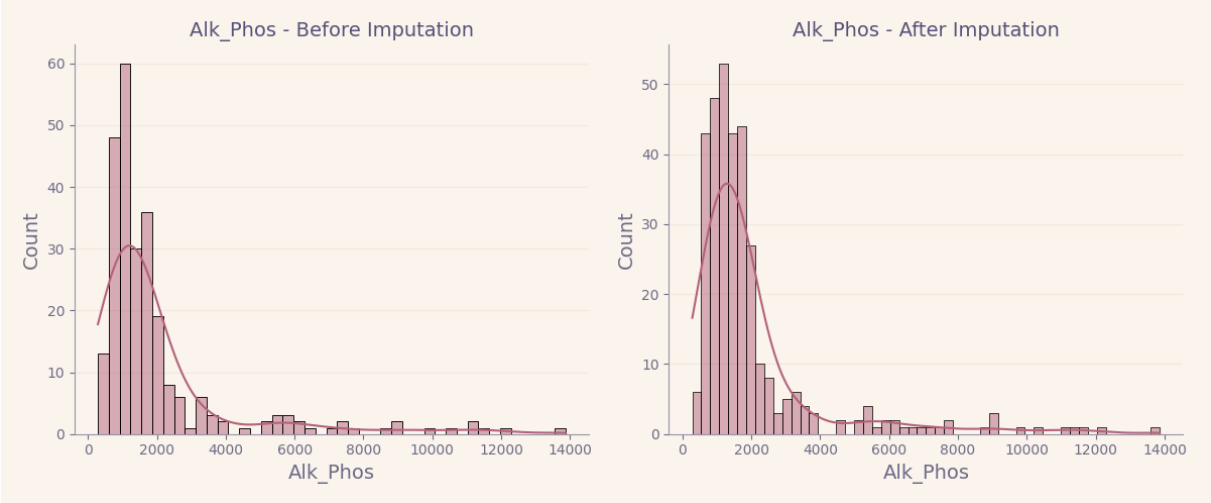


Figura 27: Histograma tractament missings variable Alk_Phos

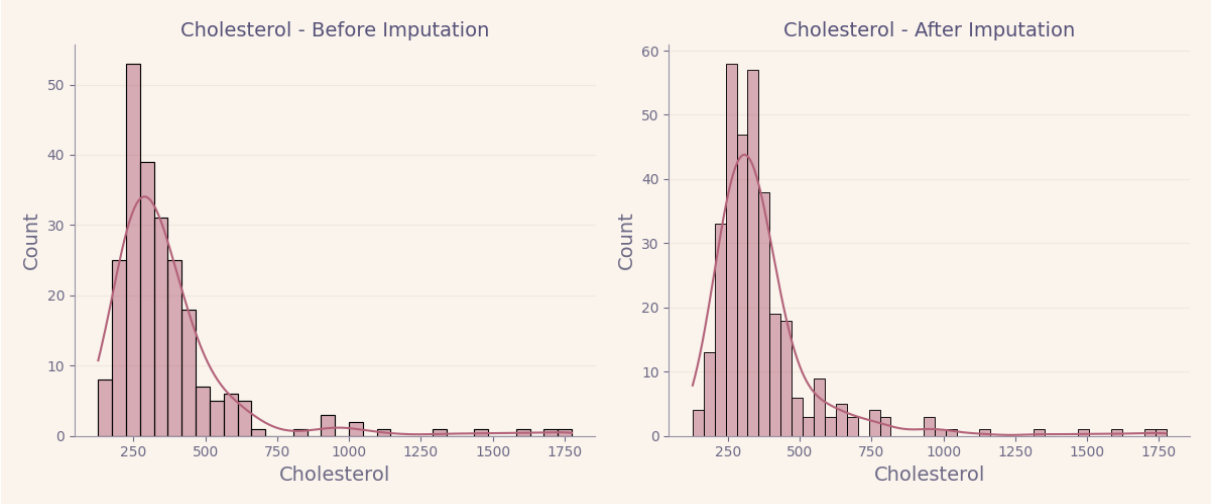


Figura 28: Histograma tractament missings variable Cholesterol

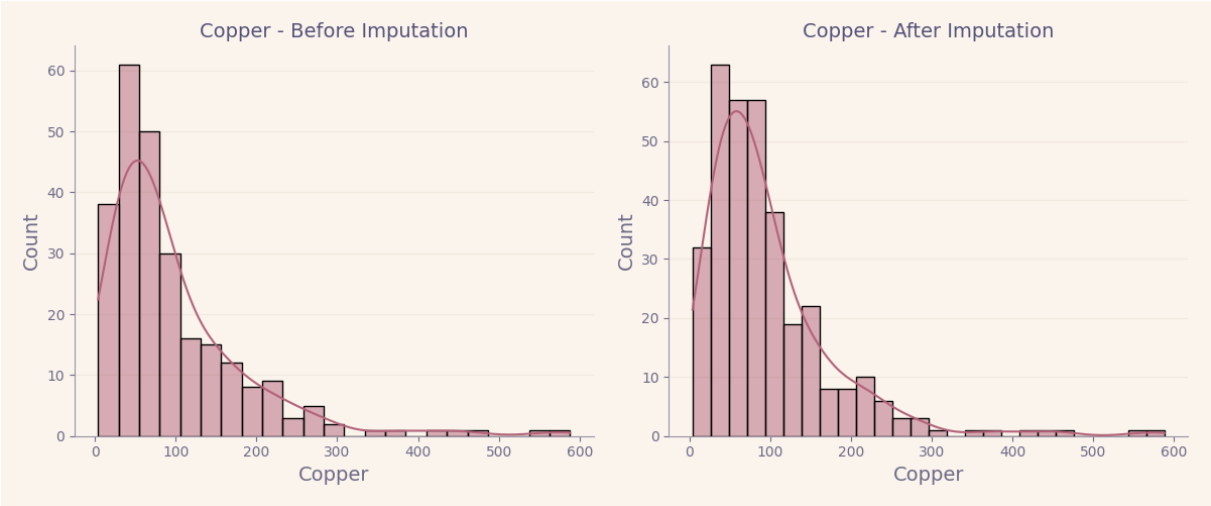


Figura 29: Countplot tractament missings variable Copper

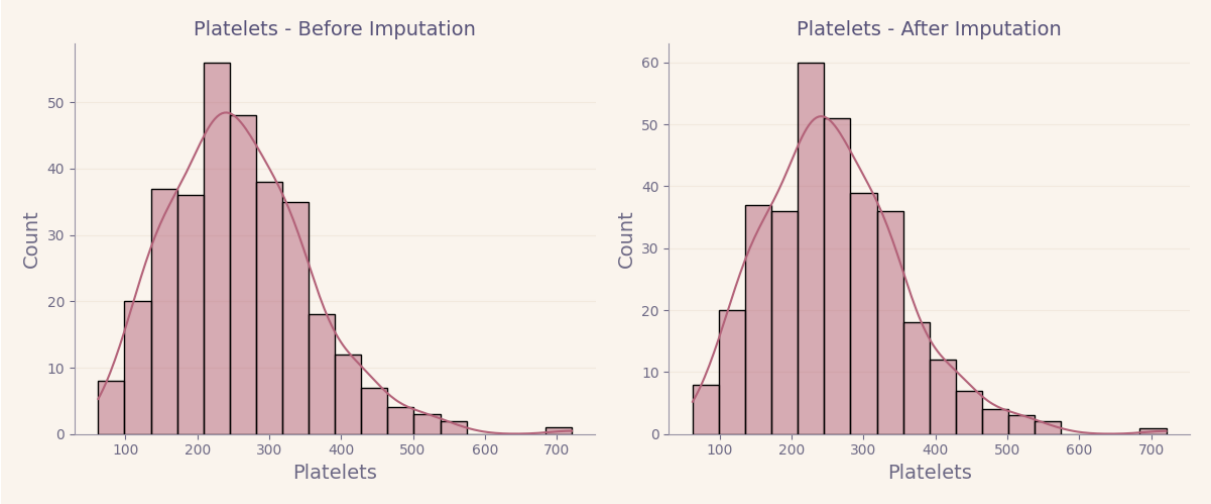


Figura 30: Countplot tractament missings variable Platelets

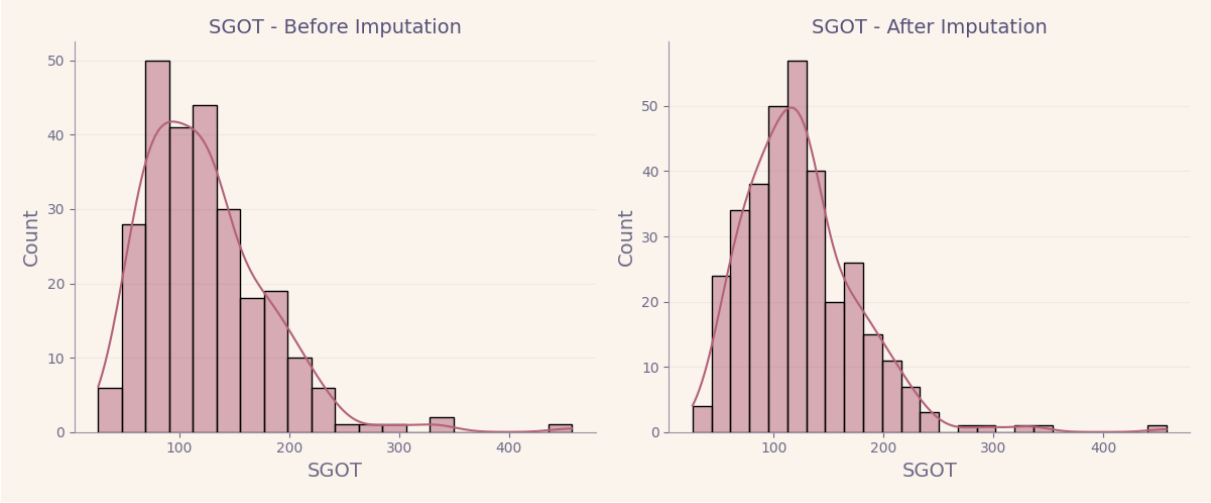


Figura 31: Countplot tractament missings variable SGOT

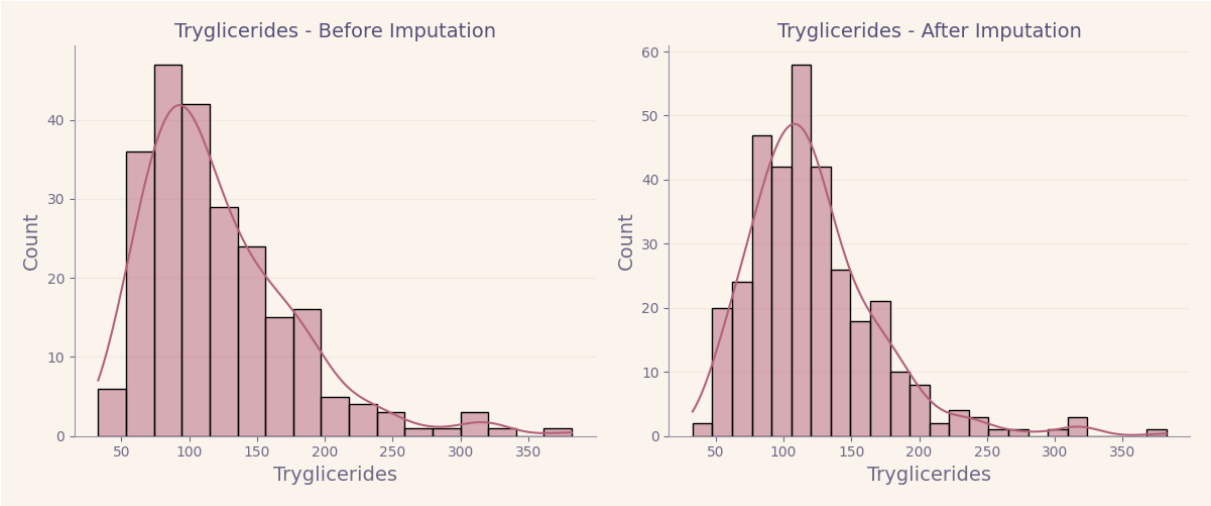


Figura 32: Countplot tractament missings variable Trygliceride

Les figures superiors mostren la imputació amb IterativeImputer usant RandomForest. Aquesta elecció, però, no s'ha fet d'una manera precipitada. S'ha realitzat una comparativa de les imputacions en la llibreta Jupyter on s'ha vist com la moda modifica totalment els histogrames de les variables categòriques, i el KNN imputava d'una manera més rudimentària que l'opció escollida.

1.4 Outliers

En aquest apartat discutirem la identificació i proposta de gestió dels outliers.

Pel que fa a la identificació, els he trobat tant de forma univariant com de forma multivariant.

- **Univariant:** De forma univariant, hem realitzat els boxplots, i vist les instàncies que estan més enllà del 1.5 IQR, més específicament les que es veuen fora dels límits (whiskers) del boxplot. En total tenim 96 otutlier univariants únics.
- **Multivariant:** De forma multivariant, he utilitzat un IsolationForest. Aquest model ha identificat un nombre de 29 outliers.

Pel que fa a la proposta de gestió, cal primer presentar uns quants punts:

- **Mida del dataset:** El nostre train té únicament 334 instàncies.
- **Models:** Casualment tots els models que s'usaran són robustos a outliers. L'únic que pot causar problemes és la convergència en models com l'SVM.
- **Coneixement del domini:** No tenim coneixement mèdic ni cap expert de la salut per dir-nos o no si els valors són correctes o possibles.
- **Risc:** Treure outliers amb tan poques mostres pot reduir la generalització i evaluació correcta de les mètriques radicalment.

Valorant la situació en la qual ens trobem i segons el meu judici, no trobo adequat eliminar aquestes instàncies. Si ens trobem amb la necessitat de fer-ho, però, ho farem.

Tot i haver-hi pres aquesta decisió, però, fem un experiment de performance traient i sense treure els outliers. On podem veure la millora de les mètriques d'avaluació en mantenir-los.

1.5 Recodificació de variables

Quant a la recodificació de variables, s'ha hagut de tractar les variables categòriques per tal de poder entrenar els diferents models. Tot i que els arbres accepten variables categòriques i numèriques, la implementació DecisionTreeClassifier només n'accepta de numèriques. Així doncs, s'ha hagut de fer diverses transformacions.

Abans d'aquestes transformacions, però, hem hagut de fer alguns canvis els quals semblen errors de codificació com ja hem dit amb anterior:

- **Canvi dtype:** S'hi ha hagut de codificar les variables segons el seu tipus real explicitat a la metadata. Així doncs, les variables numèriques que s'han hagut de codificar com a tals són: 'N_Days', 'Age', 'Bilirubin', 'Cholesterol', 'Albumin', 'Copper', 'Alk_Phos', 'SGOT', 'Tryglicerides', 'Platelets', 'Prothrombin'. Les variables categòriques que s'ha hagut de codificar com a tal són: 'Drug', 'Sex', 'Ascites', 'Hepatomegaly', 'Spiders', 'Edema', 'Stage','Status'.

- **Errors 'NaNN':** Com ja hem explicat anteriorment amb més detall, ja que aquesta categoria adherida a algunes variables sembla no tenir sentit inherent segons la metadada, l'he hagut de passar a np.nan i imputar-la de forma corresponent.

Les transformacions fetes han sigut les següents:

- **LabelEncoder:** Les variables binàries s'han tractat amb un LabelEncoder. La raó ve del fet que el model s'ha d'entrenar amb dades numèriques i, el OneHotEncoding no té gaire sentit en aquest tipus de categòriques. Per què no té sentit? Principalment pel fet que estem augmentant innecessàriament l'espai de cerca, i és redundant, ja que l'altra columna que es crea ja es pot deduir pel fet de ser complementàries. A més a més, una variable binària ja està "OneHotEncoded"per natura.

- **OneHotEncoder:** Les variables categòriques no binàries s'han tractat amb OneHotEncoder. On aquestes són Edema i Stage. Edema no presenta cap dubte, ja que és una variable categòrica no ordinal, per tant, necessita estrictament aquesta representació. Quant a Stage, tot i que té natura d'ordinal, no li posem OrdinalEncoder, ja que no tenim coneixement en què la distància entre una i l'altra fase sigui equidistant. Per exemple, si tinguéssim la categòrica estrelles d'un hotel, la codificaríem amb l'OrdinalEncoder, ja que té ordre i és "equidistant". En canvi, en el cas de Stage, no tinc el coneixement per saber si aquestes característiques presenten un ordre proporcional "de propietats". Si la possesim amb OrdinalEncoder ens la tractaria com una variable numèrica i no ho sembla de cap manera per l'anterior dit, la deixarem amb el OneHotEncoder.

1.6 Particionat del dataset

Com ja s'ha esmentat amb anterioritat, el particionat ha estat realitzat fins i tot abans de l'EDA exhaustiu (missings, outliers...) segons mesures per contenir el data leaking. Per tant:

- Els missings han estat imputats després de la partició
- Els mètodes de balanceig es realitzen després de la partició

Pel que fa a la partició, ha estat feta només en els grups train i test en una proporció de 80 i 20 respectivament, escollint un train llarg per la falta de mostres. He deixat la idea de fer una partició de validació i l'he substituït per cross-validation per les raons següents tenint en compte la dimensió de la base de dades:

- **Millor ús de les dades:** En un dataset petit, cada dada és important. L'ús de Cross-validation permet utilitzar cada isntància tant per a l'entrenament com per a la validació en múltiples folds, maximitzant l'aprofitament d'aquestes dades.
- **Més robustesa:**Els hiperparàmetres i, per tant, el rendiment del model, s'ajustaran a les mostres específiques del conjunt de validació. Amb cross-valdiation, fem la mitjana de les mètriques per aquests hiperparàmetres, fet que ajuda a tenir un model més robust i permet tenir una millor avaluació del rendiment del model.
- **Overfitting** En un dataset petit, hi ha un risc més alt d'overfitting perquè el model pot aprendre a funcionar bé en les mostres específiques del conjunt de validació. El cross validation ajuda a mitigar aquest risc mitjançant l'avaluació del model en diferents subconjunts de les dades.

Pel que fa a la metodologia de particionat, s'ha fet un estudi comparant l'ús del particionat estàndard, contra l'ús d'estratificació per la variable objectiu.

La taula resultant ha sigut la següent:

Status Category	Overall %	Stratified %	Random %	Strat. Error %	Rand. Error %
C	55.50	55.95	52.38	0.81	-5.62
CL	5.98	5.95	4.76	-0.48	-20.38
D	38.52	38.10	42.86	-1.09	11.27

Veient els resultats, com és obvi, estratificarem segons la categoria Status per a ajustar-nos a les proporcions de la població general en el train.

2 Preparació de variables

2.1 Normalització de variables

En aquest apartat discutirem les opcions que tenim de cara a normalitzar les nostres variables, així com quina s'ha utilitzat de primera mà abans de realitzar cap comparació.

En primer lloc, cal destacar que els models basats en arbres de decisió que es faran en els apartats següents no són afectats per l'escala de les variables, on intentarem no utilitzar aquests mètodes per interpretar la informació que ens donen.

Pel que fa al KNN i SVM, cal discutir quin mètode utilitzarem per a escalar les dades. Principalment, realitzarem la comparativa entre MinMaxScaler (1) i StandardScaler (2). Ara, però, abans de dur a terme la comparativa directament sobre el rendiment dels models, decidirem quin usar en primera mà segons els punts següents.

Mètode	Pros	Contres
1	Bona interpretabilitat, ja que deixa les dades en una escala [0,1]	Molt sensible a outliers (els quals no hem tret de primera mà), modifica la distribució de les dades si no s'aproximen a una uniforme o té molts zeros, no contempla l'arribada de pacients fora del rang d'escalat.
2	No canvia les distribucions, menys sensible a outliers.	Menor interpretabilitat, pot canviar les distàncies relatives entre columnes.

Tot i que farem un experiment comparant els dos mètodes a l'apartat 3, començarem utilitzant el StandardScaler, ja que no tractarem els outliers i sembla més estable a primera vista.

Els resultats són els següents:

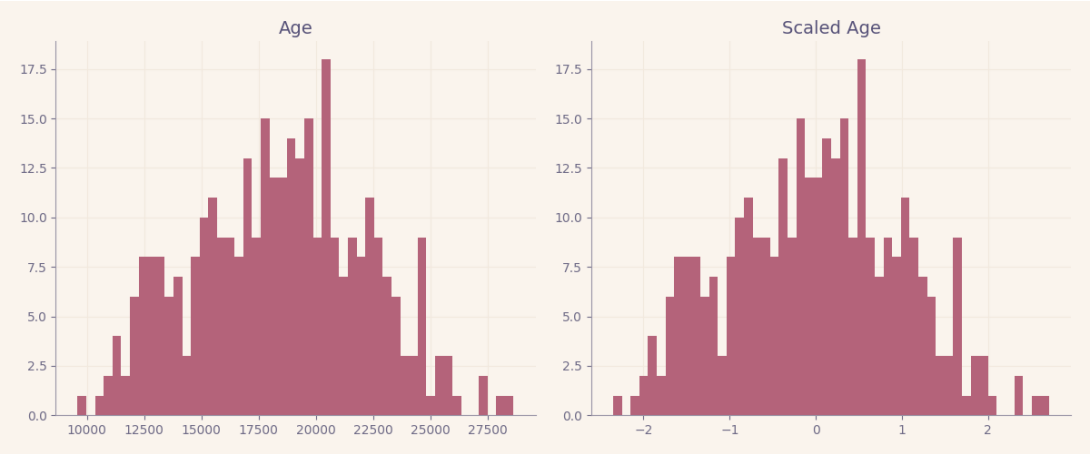


Figura 33: Histograma post escalat variable Age

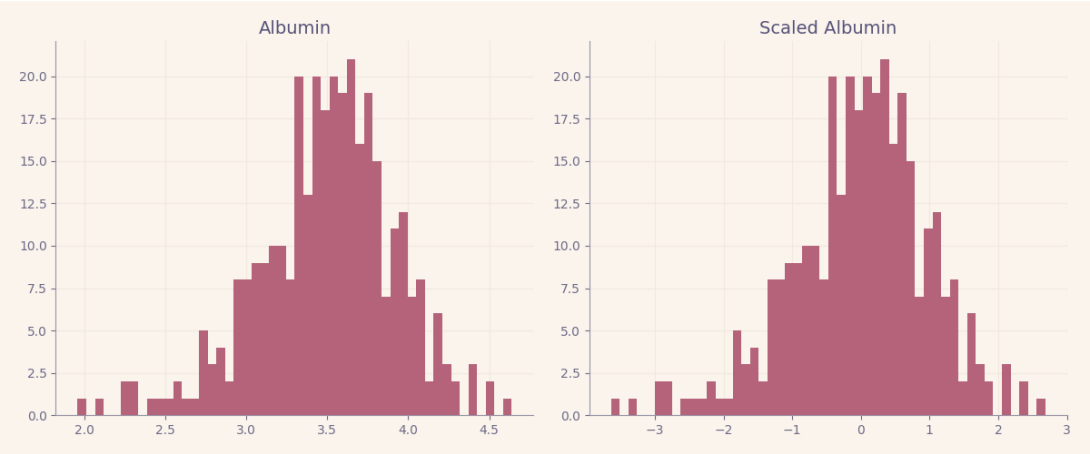


Figura 34: Histograma post escalat variable Albumin

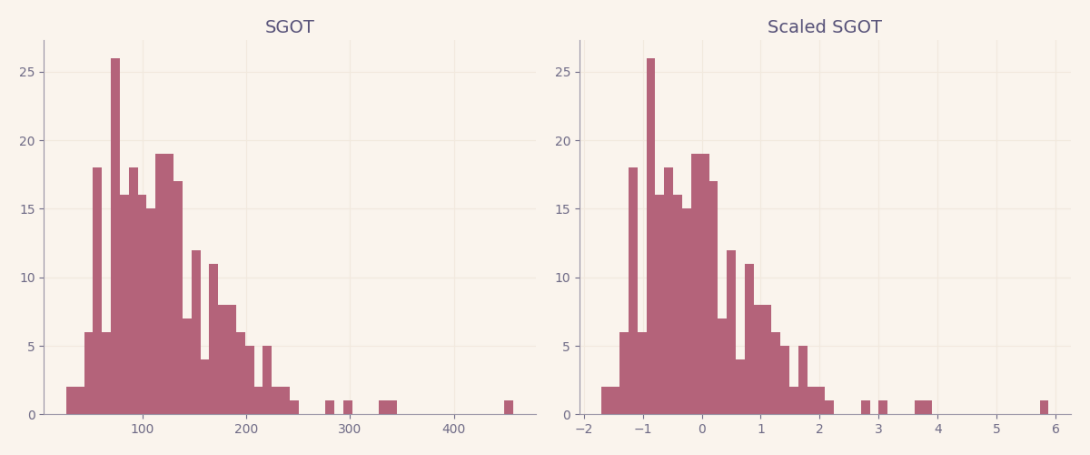


Figura 35: Histograma post escalat variable SGOT

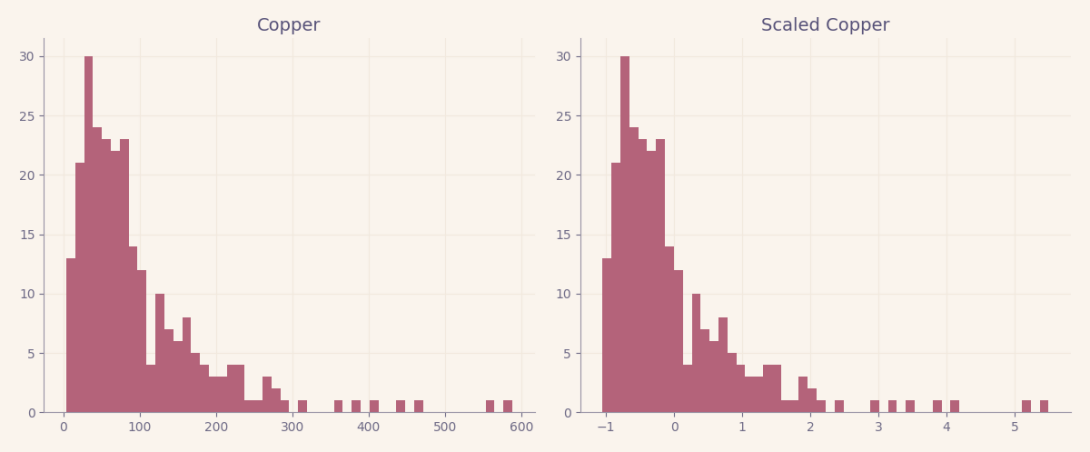


Figura 36: Histograma post escalat variable Copper

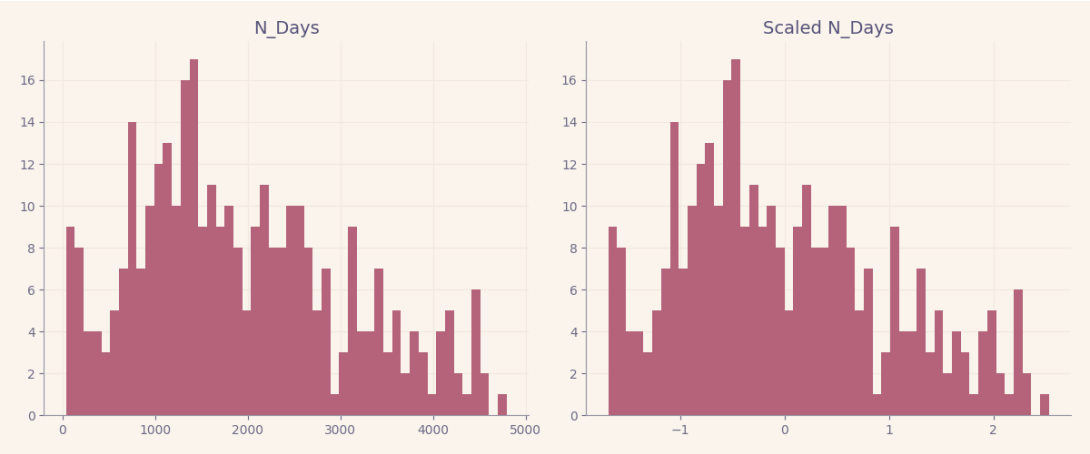


Figura 37: Histograma post escalat variable N_Days

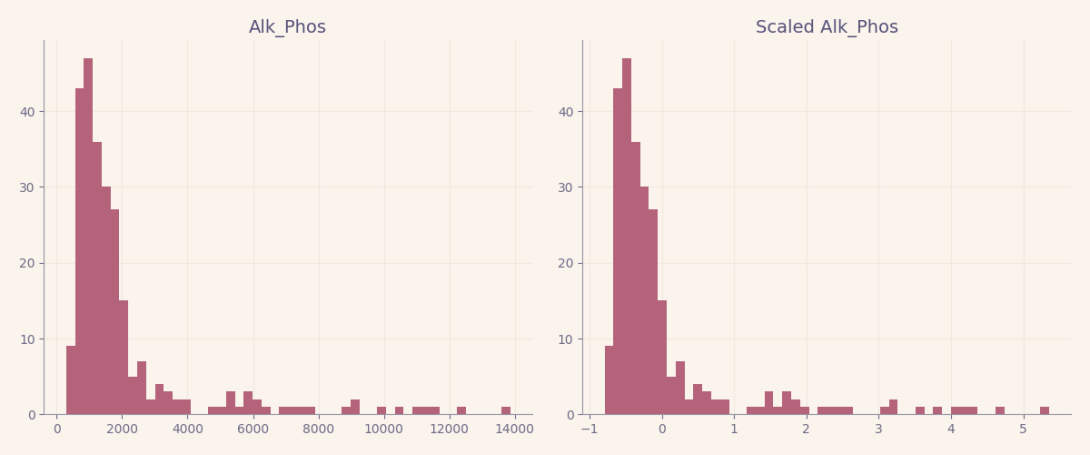


Figura 38: Histograma post escalat variable Alk_Phos

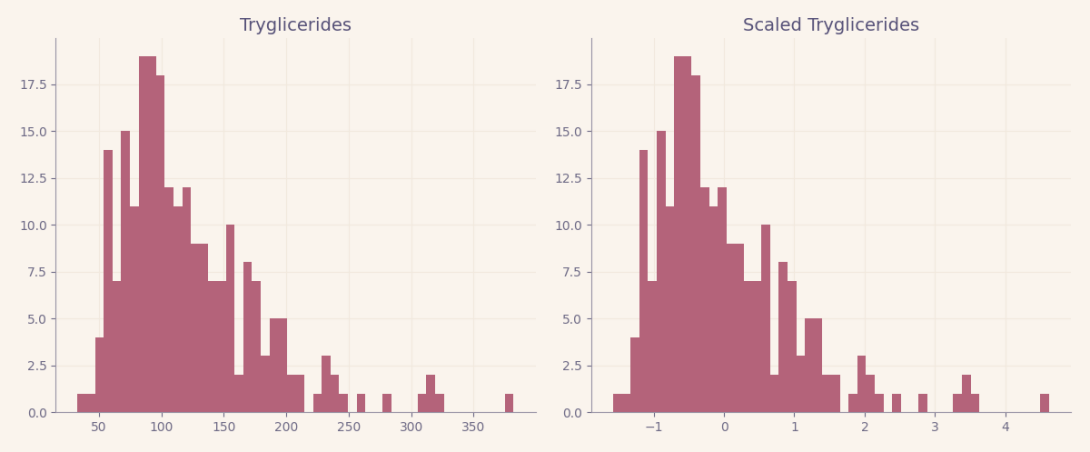


Figura 39: Histograma post escalat variable Triglicerides

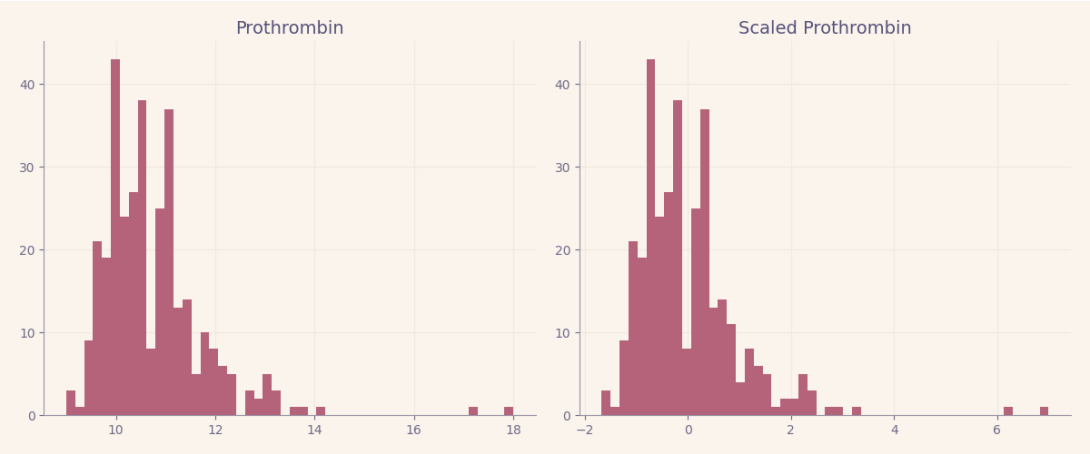


Figura 40: Histograma post escalat variable Prothrombin

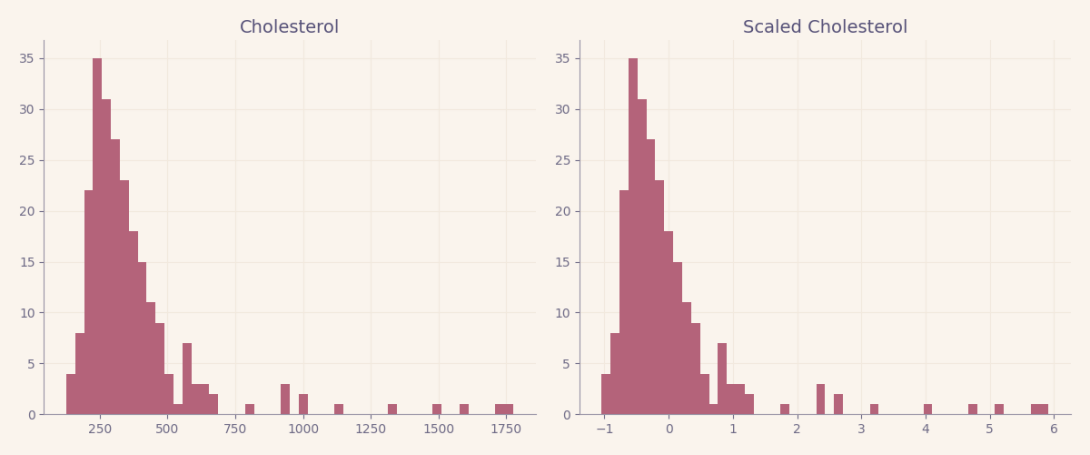


Figura 41: Histograma post escalat variable Cholesterol

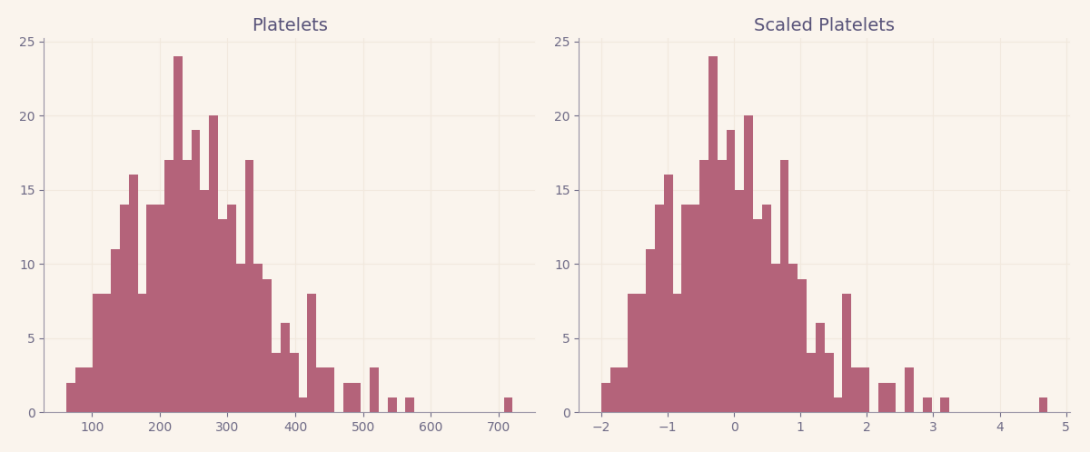


Figura 42: Histograma post escalat variable Platelets

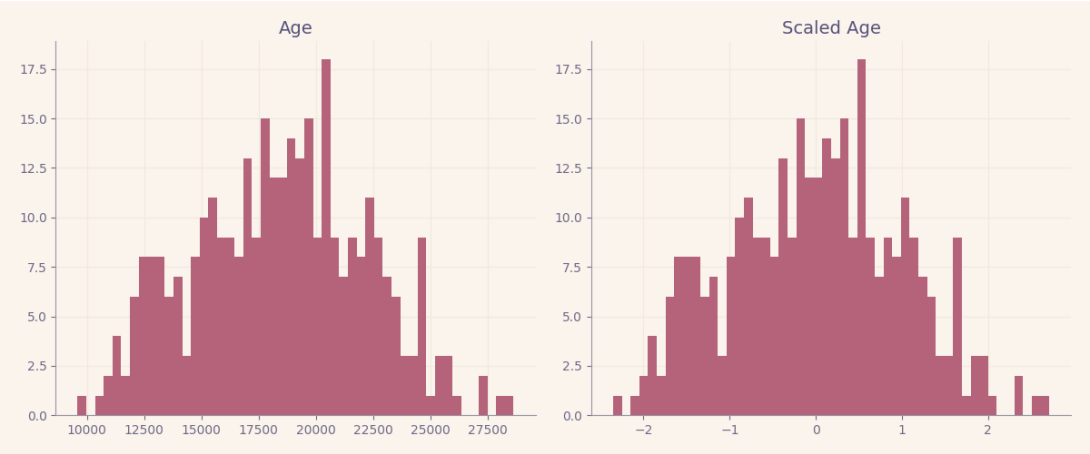


Figura 43: Histograma post escalat variable Bilirubin

2.2 Anàlisi de correlacions entre variables númeriques

Abans de seleccionar variables i entrenar models, cal analitzar les correlacions (Pearson per relacions lineals i Spearman per les relacions no lineals) entre les variables numèriques que conformen la nostra base de dades:

Correlació de Pearson:

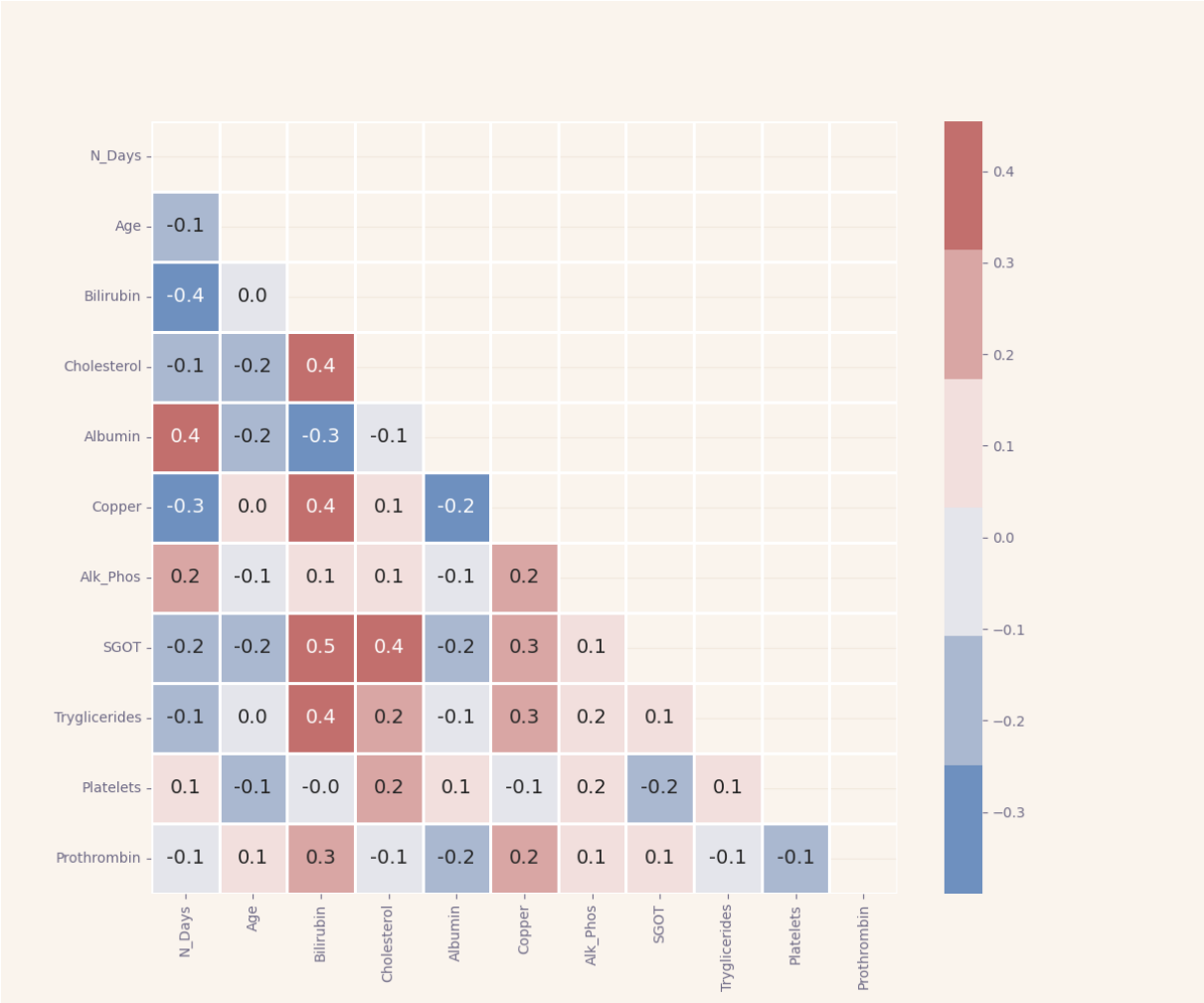


Figura 44: Heatmap correlació Pearson variables numèriques

Com podem veure, no trobem cap relació lineal forta, on en podem trobar alguna moderada (0.4).

Correlació de Spearman:

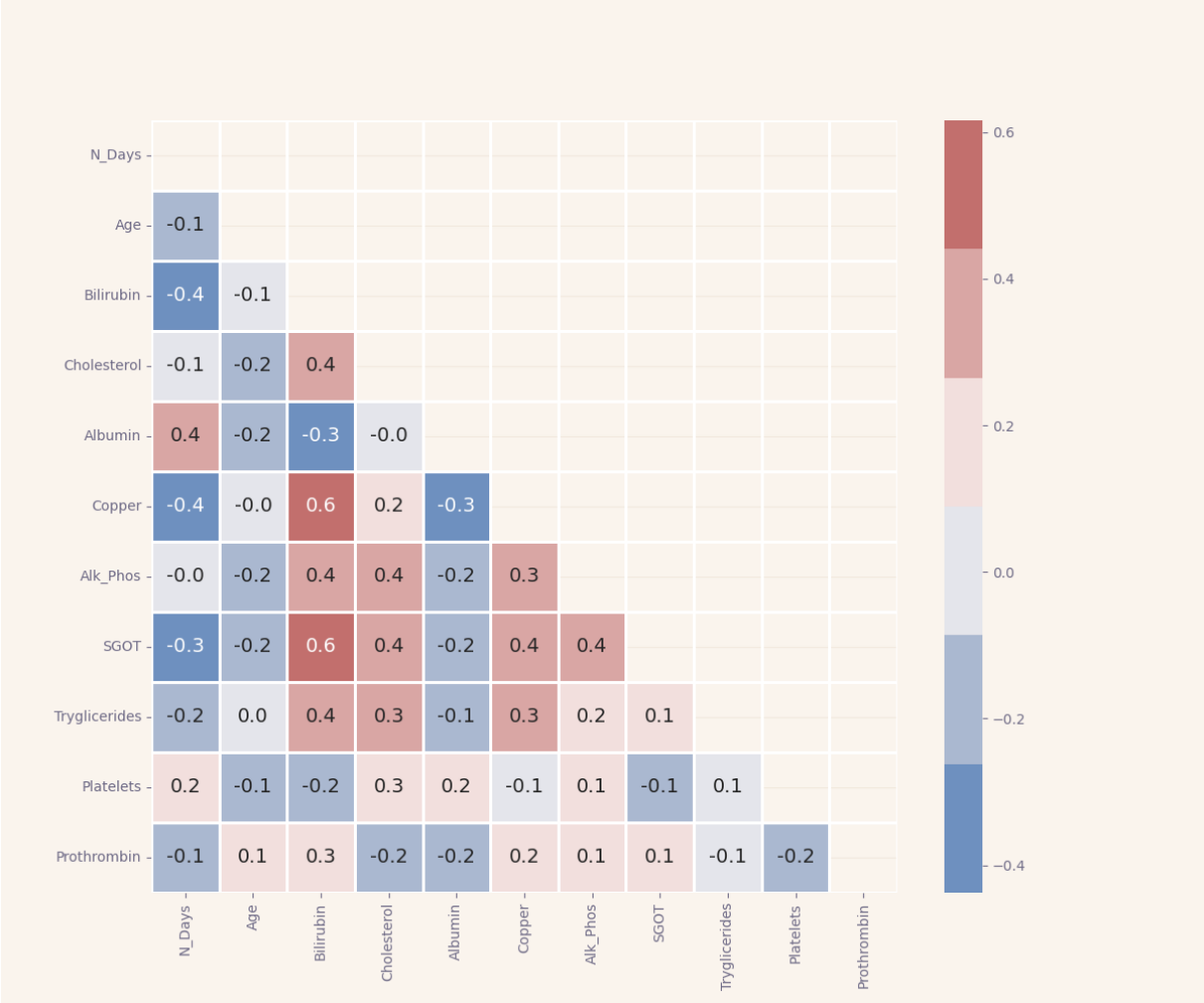


Figura 45: Heatmap correlació Spearman variables numèriques

Aqui podem veure algunes relacions més fortes, aquestes no-lineals. Destaca la relacio de Copper i SGOT amb Bilirubin.

Quant a filtrar variables per la seva forta correlació amb altres, no crec que es pugui aplicar a les nostres dades, ja que ens trobem amb vincles moderats com a màxim.

2.3 Anàlisi de variables categóriques i variable objectiu

En aquest apartat veurem la influència que poden tenir diverses variables categòriques a l'hora de predir la variable objectiu 'Status' així com la seva possible influència.

Per realitzar aquest procés s'ha optat per realitzar countplots normalitzats (plots de freqüència) per grups.

El resultat ha sigut el següent:

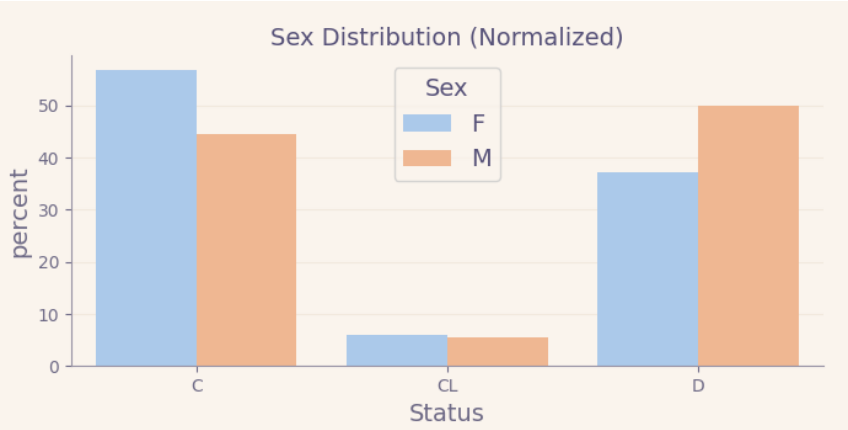


Figura 46: Countplot normalitzat per classe variable Sex

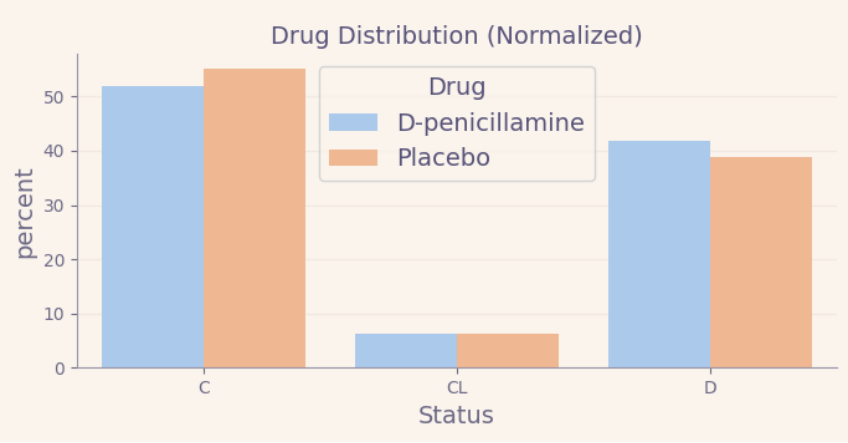


Figura 47: Countplot normalitzat per classe variable Drug

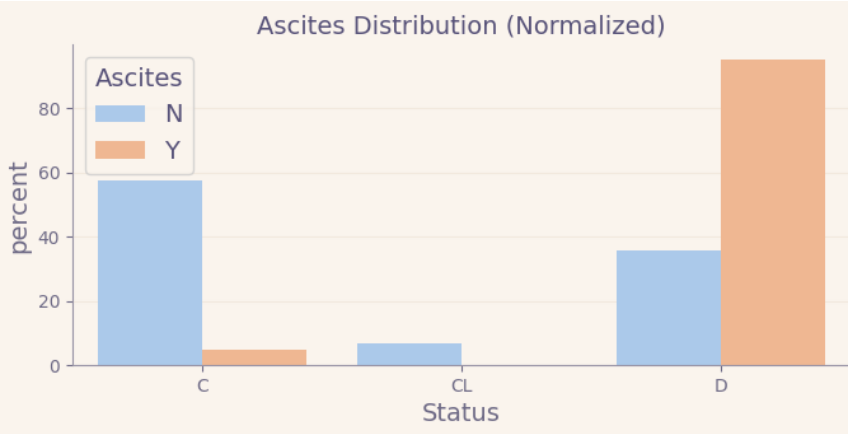


Figura 48: Countplot normalitzat per classe variable Ascites

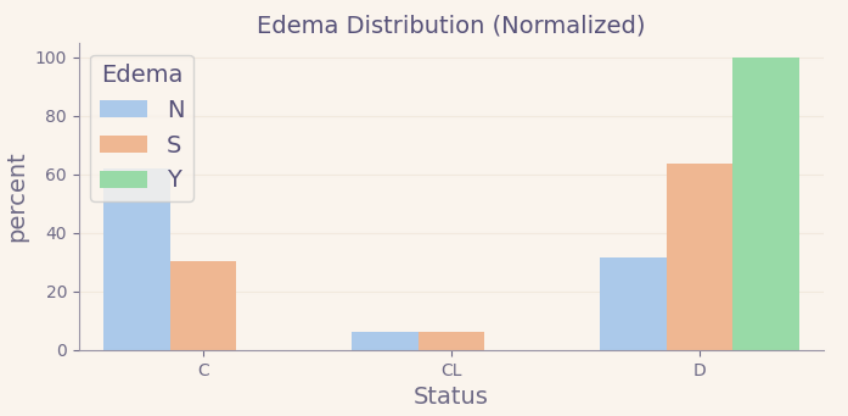


Figura 49: Countplot normalitzat per classe variable Edema

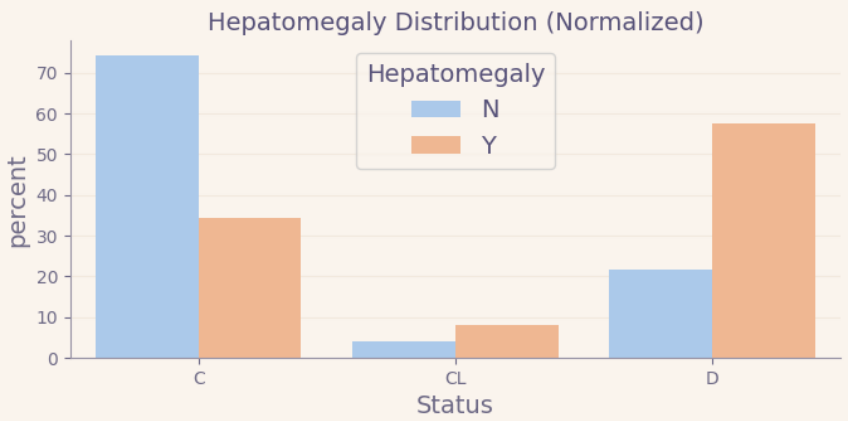


Figura 50: Countplot normalitzat per classe variable Hepatomegaly

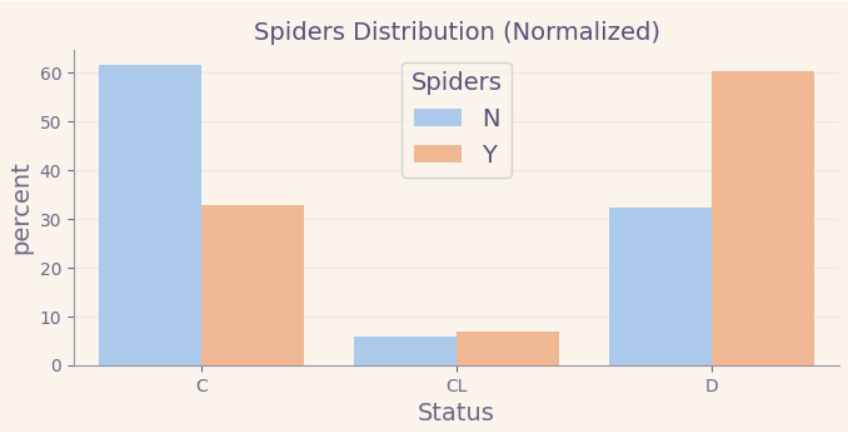


Figura 51: Countplot normalitzat per classe variable Spiders

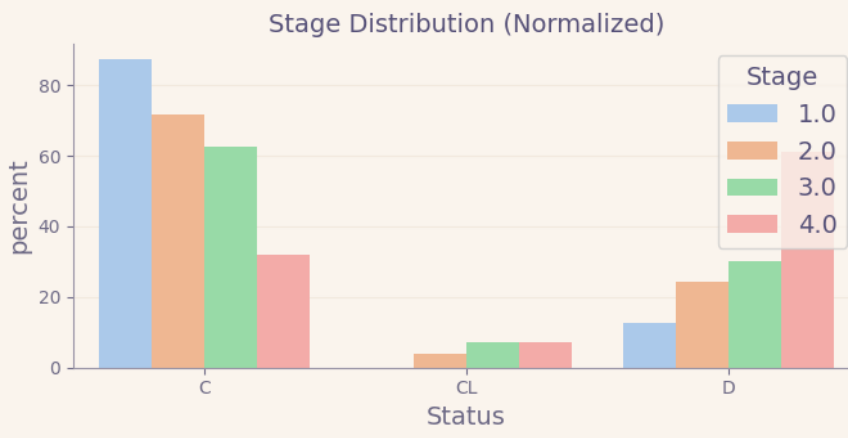


Figura 52: Countplot normalitzat per classe variable Stage

A primera vista, podem extreure algunes conclusions:

- **Status:** Amb tres categories, la més freqüent és la C (viu sense trasplantament), que representa el 55% de les ocurrències. Aquest desequilibri és notable i cal tenir-ho en compte.
- **Drug:** No veiem una diferència significativa de la influència del tractament respecte al 'Status'.
- **Sex:** Podem observar una possible major taxa de mortalitat en homes i una major taxa de supervivència sense trasplantament per dones, tot i que la diferència és petita. Pot ser una conseqüència de les poques instàncies del dataset.
- **Ascites:** Veiem clarament que la gent que pateix Ascites tendeix més a morir, mentre que la gent que no tendeix més a curar-se o a curar-se amb trasplantament.
- **Edema:** Veiem que la gent que té Edema mor en gran proporció, seguit per les persones que prenen diürètics per controlar-lo i últimament per la gent que no té Edema. Fora de les morts, veiem que entre la gent que no té Edema i els que prenen diürètics només es diferencien a l'hora de recuperació sense trasplantament, on els que no en tenen sembla que són més presents.

- **Hepatomegaly:**

  Pel que fa a aquesta variable, veiem que la gent que pateix aquesta condició té més tendència a morir o a rebre trasplantament, en comparació amb la gent que no. Mentre que la gent que no la té, tendeix més a sobreviure sense trasplant.

- **Spiders:**

  En la variable Spiders, podem veure que la gent que pateix aquesta condició mor més i li fan més trasplantaments que la que no, mentre que la gent que no la pateix sobreviu més sense trasplantament.

- **Stage:**

  Pel que fa a la variable Stage, podem apreciar, com és lògic, el descens de freqüència de supervivència sense trasplantament quan més avançada és la fase. De la mateixa manera, veiem com en la gent que mor, hi ha més freqüència en les fases altes de forma ascendent.

2.4 Anàlisi de variables numèriques i variable objectiu

En aquest apartat veurem la influència que poden tenir diverses variables numèriques a l'hora de predir la variable objectiu 'Status' així com la seva possible influència. Les conclusions, a més dels boxplots, comentaràn les distribucions per categoria que hem vist ja a l'apartat d'anàlisi univariant:

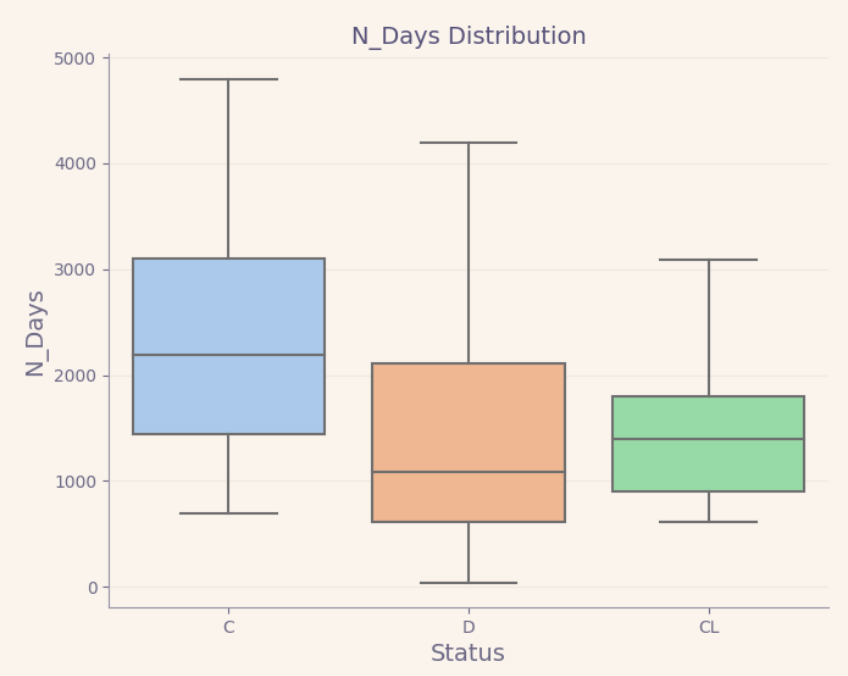


Figura 53: Boxplot per classe variable N_days

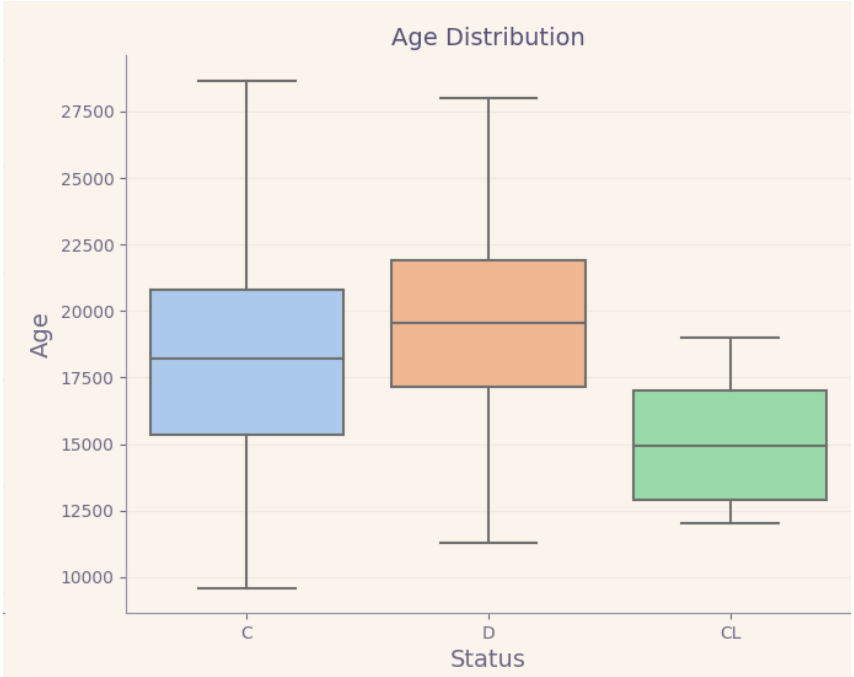


Figura 54: Boxplot per classe variable Age

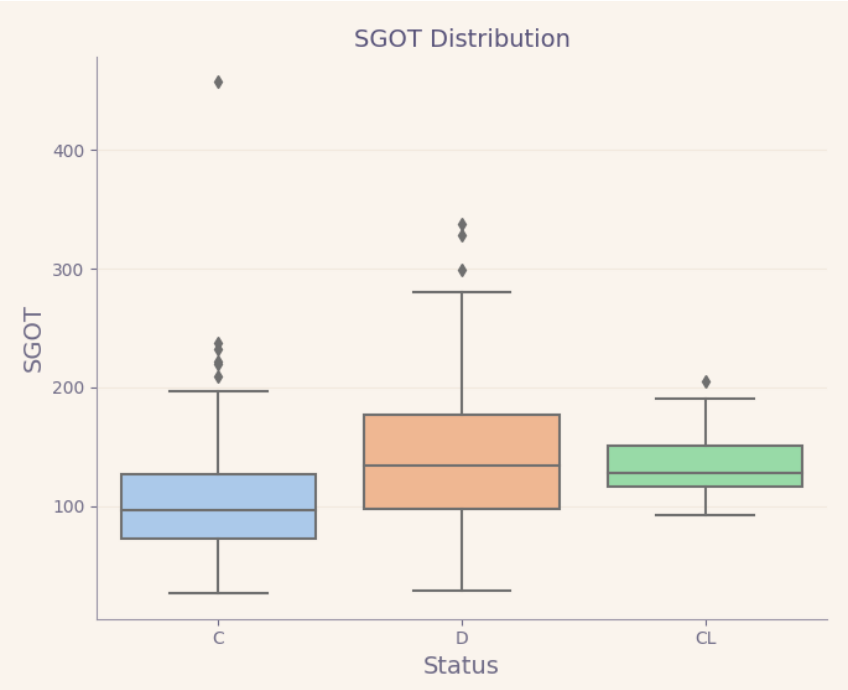


Figura 55: Boxplot per classe variable SGOT

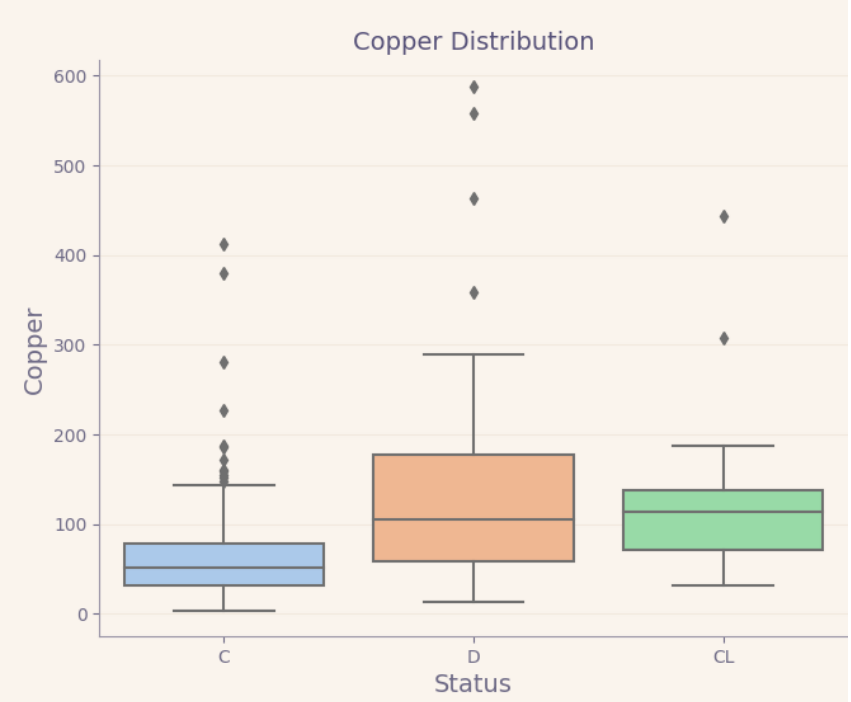


Figura 56: Boxplot per classe variable Copper

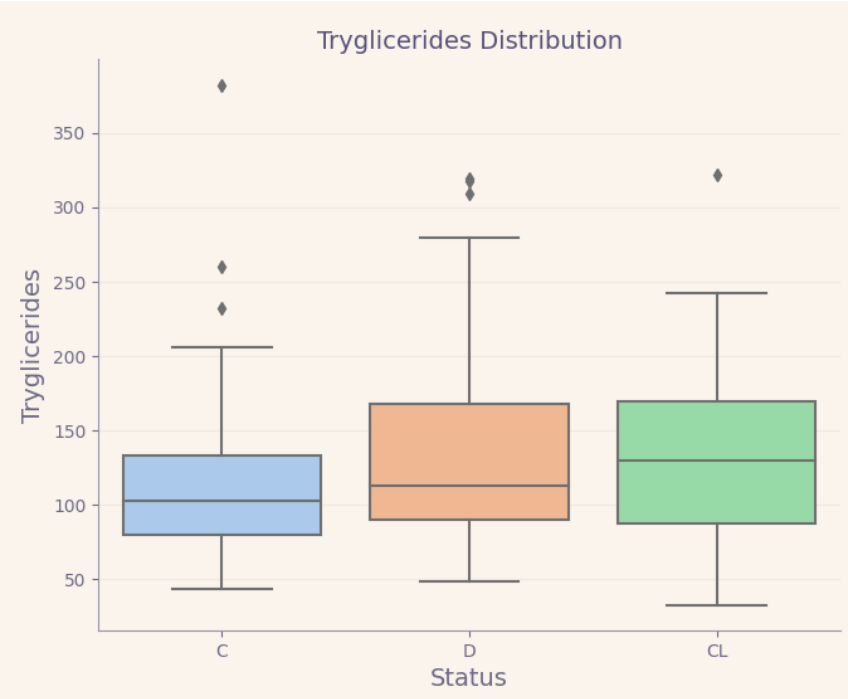


Figura 57: Boxplot per classe variable Trygliceride

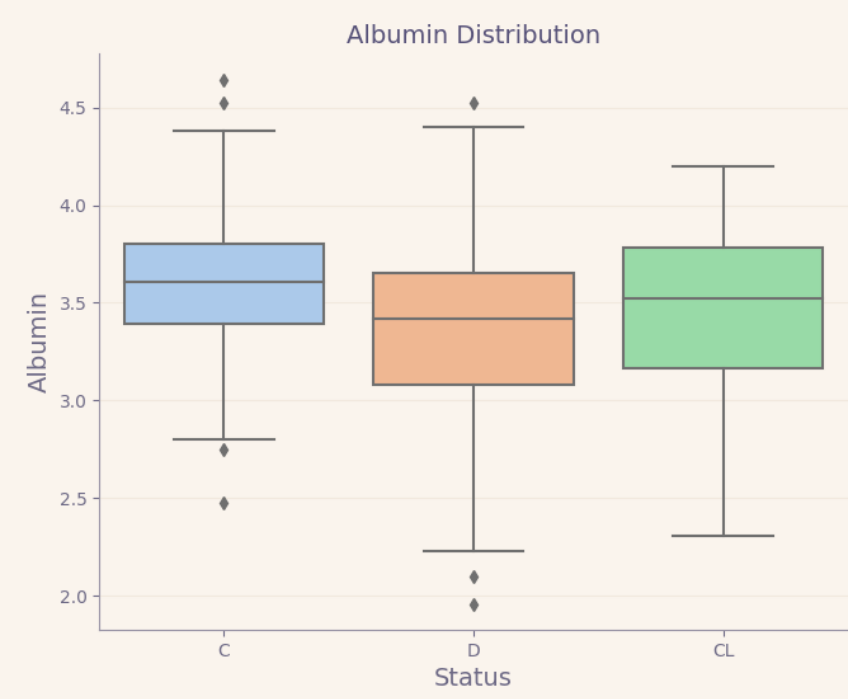


Figura 58: Boxplot per classe variable Albumin

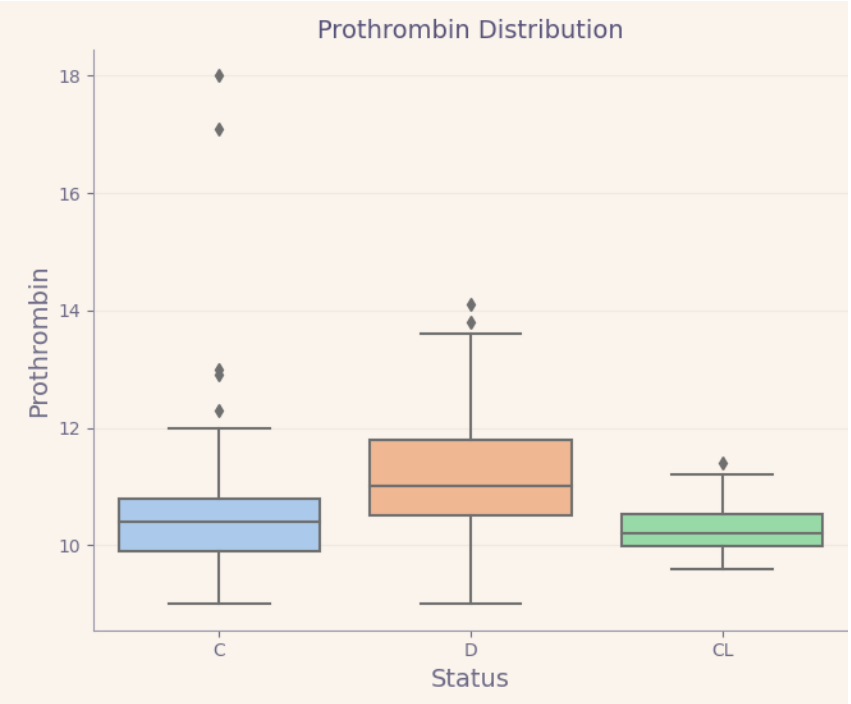


Figura 59: Boxplot per classe variable Prothrombin

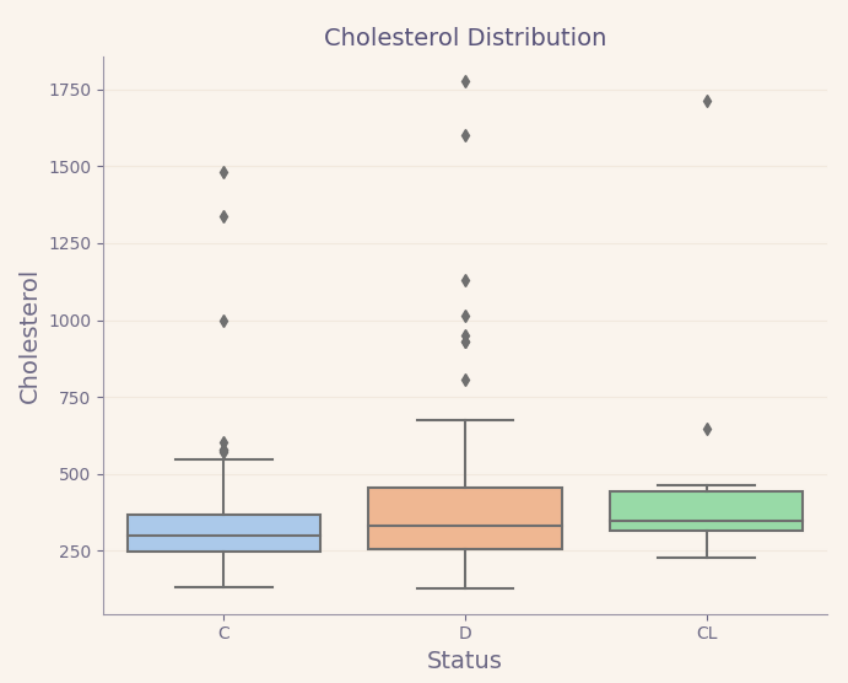


Figura 60: Boxplot per classe variable Cholesterol

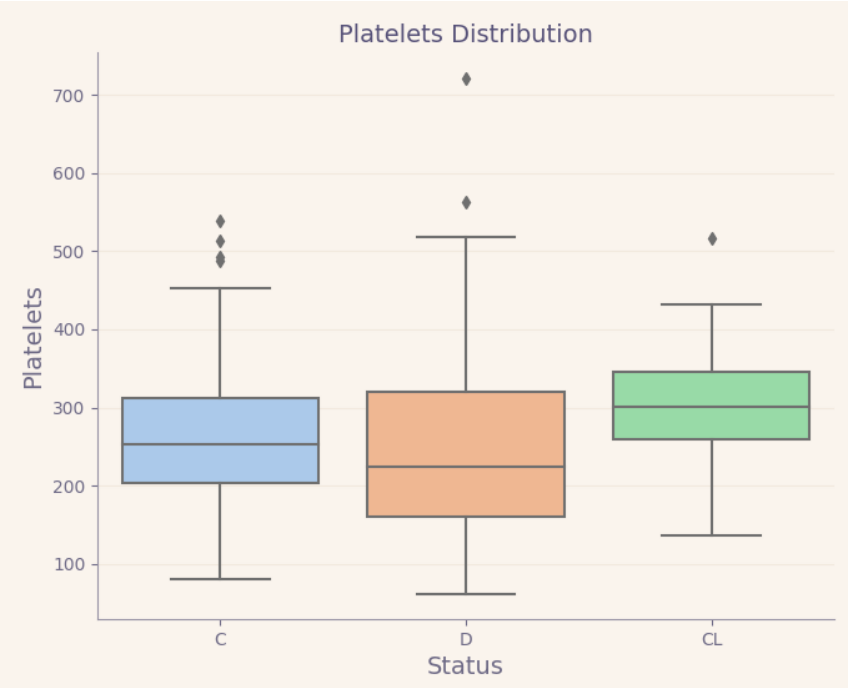


Figura 61: Boxplot per classe variable Platelets

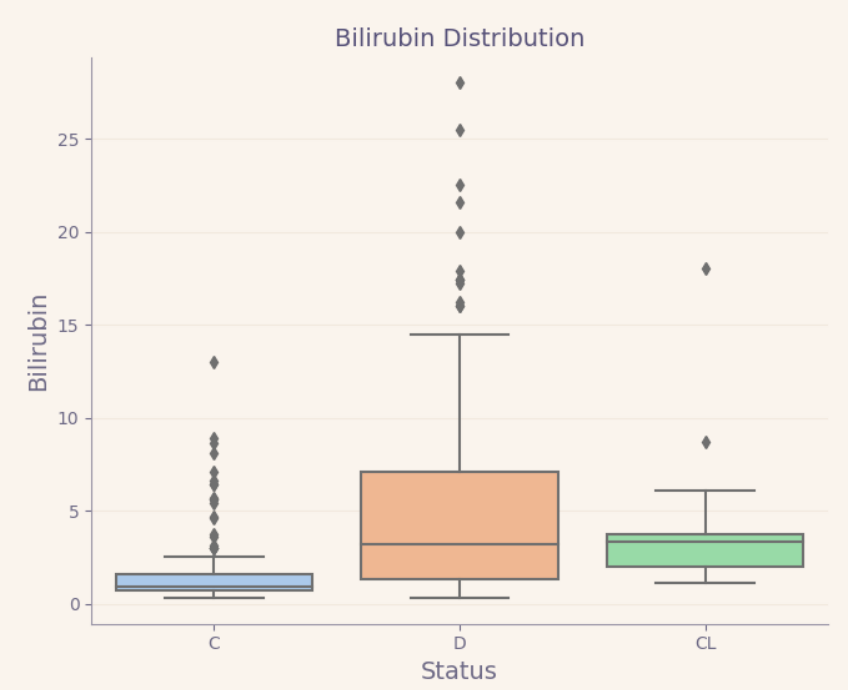


Figura 62: Boxplot per classe variable Bilirubin

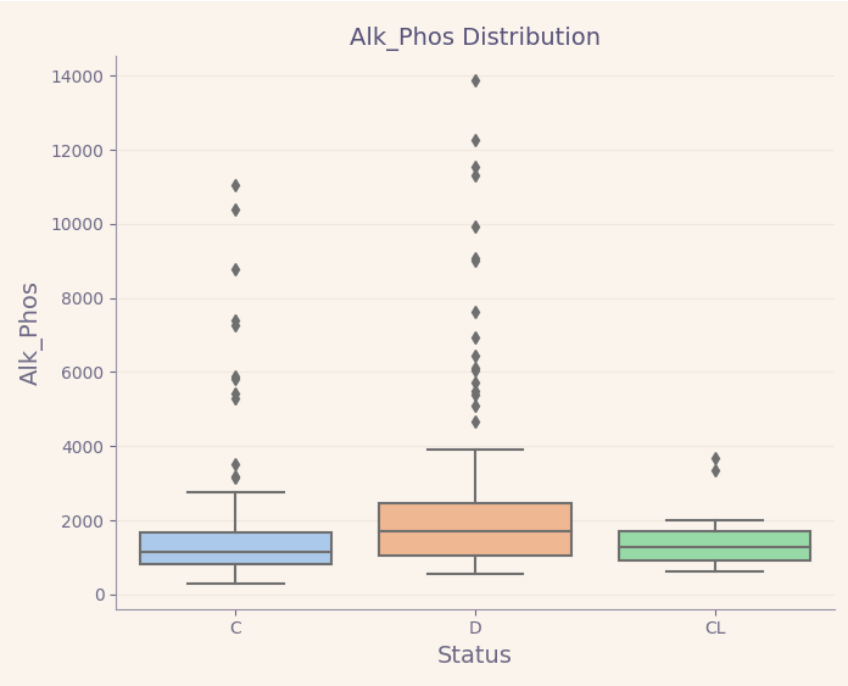


Figura 63: Boxplot per classe variable Alk_Phos

Basant-se en les discussions anteriors, sembla que els pacients de la categoria D són molt diferents que els pacients de la categoria C i CL. Seria més fàcil diferenciar D de C i CL. La part qüestionadora és determinar amb precisió els pacients de classe CL i diferenciar-los dels pacients de classe C

- **Age:** Les dades mostren que la cirrosi es produeix principalment en persones de més de quaranta anys, amb la mort principalment en persones de més edat, i on la majoria d'elles aconsegueixen el seu trasplantament de fetge abans dels cinquanta anys.
- **Bilirubine:** Com a més bilirubina en sang, sembla que hi ha més complicacions amb el fetge. Les persones que moren molt probablement tenien danys greus indicats pel nivell més alt de bilirubina. Els supervivents sense trasplantament tenen menys nivells de bilirubina que indiquen que la majoria d'ells podien estar en control (no molt malament). Les persones que necessiten trasplantament de fetge també mostren nivells més alts.
- **Albumin:** Sembla que la producció d'albúmina disminueix a causa de problemes amb el fetge. Això es pot observar en els pacients que no sobreviuen a mesura que els nivells d'albúmina disminueixen, en comparació amb els supervivents sense trasplantament o les persones amb trasplantament.
- **Copper:** Podem veure l'acumulació de coure en pacients morts o amb trasplantament. Fet que es pot afirmar mirant la seva mediana.
- **Alk_Phos:** Podem analitzar com la mediana és bastant alta en pacients D i CL.
- **SGOT:** Els nivells de SGOT poden augmentar segons la gravetat. La mediana és bastant alta en pacients D i CL

- **Tryglicerides:** Els nivells de triglicèrids poden augmentar a causa de la gravetat en la cirrosis. La mediana és lleugerament alta en pacients amb D i CL
- **Platelets:** El recompte de plaquetes sembla disminuïr a mesura que avança la cirrosi És significativament menor en les persones que no han sobreviscut.
- **Prothrombin:** Sembla que el temps de Protrombina, es pot prolongar en la cirrosi. Una altra vegada, significativament alt en les persones que no van sobreviure.

Basant-se en les discussions anteriors, sembla que els pacients de la categoria D són molt diferents que els pacients de la categoria C i CL. El repte realment és determinar amb precisió els pacients de classe CL i diferenciar-los dels pacients de classe C.

2.5 Eliminació de variables redundants o sorolloses

Pel que fa a l'eliminació de variables per reduir l'espai de cerca ens podem basar en dos criteris dels vists anteriorment:

- **Redundants:** Les variables redundants les considerarem com les que estan amplialment correlacionades amb algunes de les nostres variables i, per tant, no aporten informació nova. Com hem vist en l'anàlisi de correlacions, però, no ens trobem amb cap situació crítica en aquest aspecte.
- **Sorolloses:** Les variables sorolloses, però seran aquelles que no afegeixen informació valuosa a les nostres prediccions sinó que n'augmenten l'espai de cerca. En aquest cas ens hem trobat amb la columna inicial 'ID' que ens identificava els pacients, i potencialment amb la columna 'Drug' segons l'anàlisi anterior.

Tot i que amb les explicacions recalcades només podem filtrar aquestes dues variables, en un futur provarem mètodes com la selecció per importància en arbres de decisió, o la selecció recursiva.

A més a més, cal recalcar que el que més ens ajudaria de cara a seleccionar variables seria el coneixement de domini i, per tant, un expert.

2.6 Addició de variables :

Tot i que tenim bastants variables, he tingut la sensació que podem explotar la informació disponible per a poder crear noves variables. On per a confirmar la seva utilitat, durem a terme també l'anàlisi per classe.

En primer lloc, vaig tenir la intenció de crear una variable que aproximés la gravetat d'un pacient. Ja que tenim 4 variables que expressen la presència de diferents patologies derivades de les cirrosis, aquesta variable **N_conditions** expressa el nombre de patologies derivades de les cirrosis que presenta el pacient.

En segon lloc, vaig pensar que tot i que tenim els nivells de diferents indicadors en format numèric (eg. Bilirubina), la medicina estableix uns intervals o llindars per a denominar noves condicions, així com una escala discreta de quan podem tenir problemes amb un cert indicador de salut. On aquests intervals o llindars són de coneixement mèdic general. D'aquesta idea van sortir 5 variables noves:

- **normal_bilirubin:** Aquesta nova variable expressa la normalitat en els nivells de bilirubina d'una persona.
- **low_albumin:** Ja que hem expressat que nivells baixos d'albúmina tenen lloc en pacients que han mort, els intentarem expressar amb aquesta variable.
- **normal_copper** Aquesta variable addicional inclourà la gent dins de l'interval saludable de coure.
- **high_alk_phos:** Com hem observat abans, nivells alts d'aquest fosfat poden reflectar la gravetat d'un pacient.
- **thrombocytophenia** Aquest estat mèdic indica la falta de plaquetes.

En primera instància també vaig crear la variable edat en dies per una major interpretabilitat, però més tard vaig veure que s'estava sacrificant precisió a canvi de no voler fer la conversió més tard.

La primera variable n_conditions té les següent freqüència per classe:

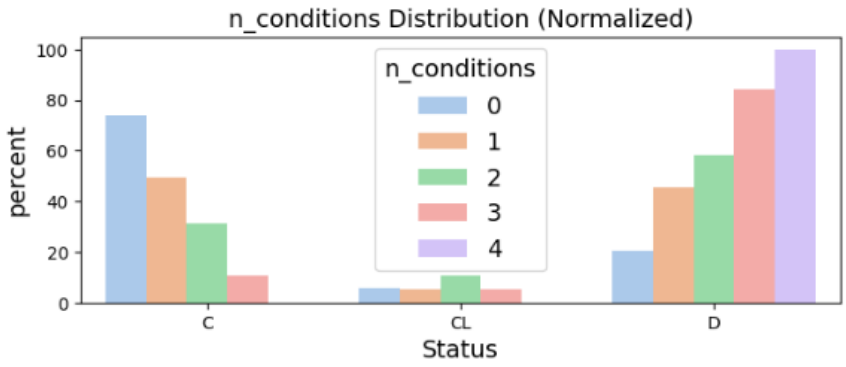


Figura 64: Countplot normalitzat per classe variable n_conditions

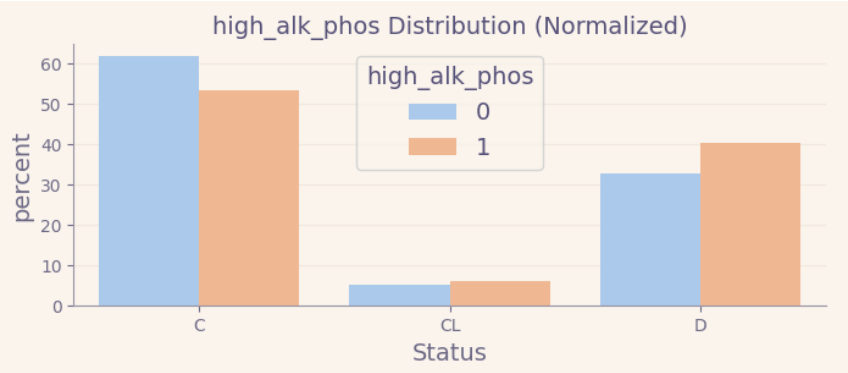


Figura 65: Countplot normalitzat per classe variable high_alk_phos

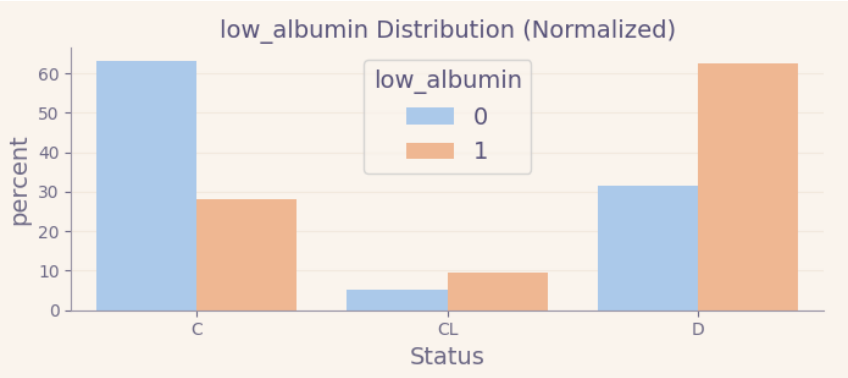


Figura 66: Countplot normalitzat per classe variable low_albumin

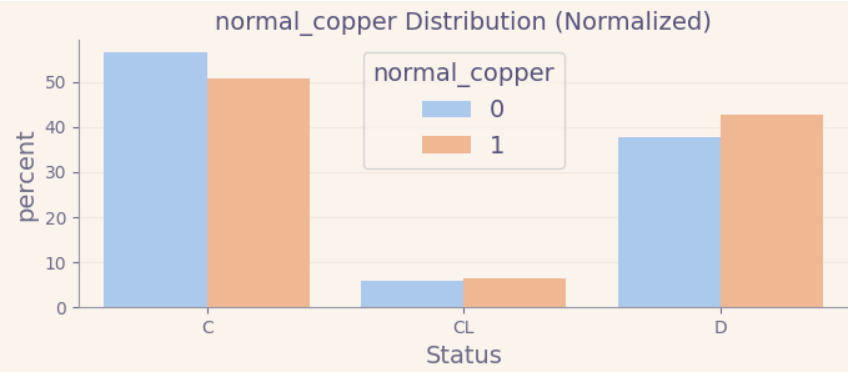


Figura 67: Countplot normalitzat per classe variable normal_copper

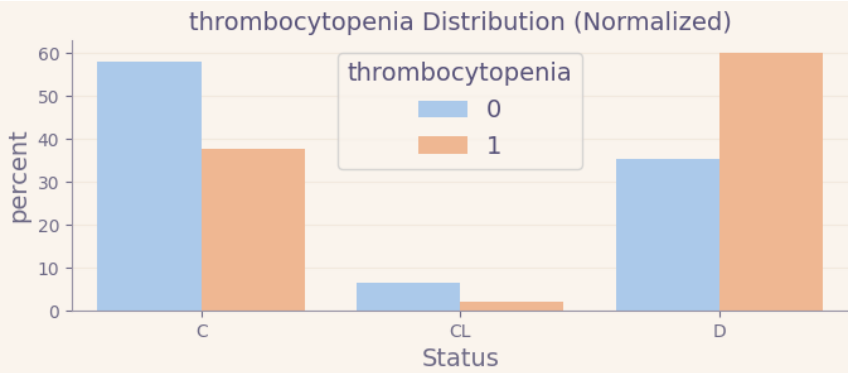


Figura 68: Countplot normalitzat per classe variable thrombocytophenia

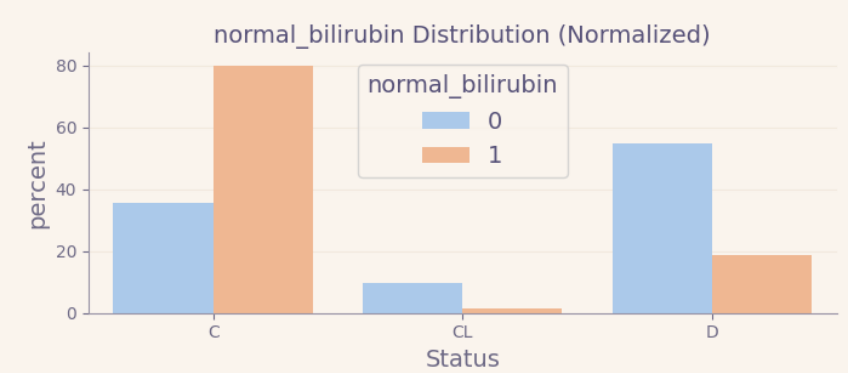


Figura 69: Countplot normalitzat per classe variable normal_bilirubin

Sembla que trobem variables bastant significatives exceptuant potser alk_Phos i Copper. Cal destacar especialment el bon comportament de n_conditions.

Pel que fa a la codificació, sols caldrà codificar n_conditions, on tindrem un OrdinalEncoder per la seva natura "equidistant", ja que expressem un recompte.

2.7 Estudi de dimensionalitat amb PCA

En aquest apartat interpretarem els resultats del PCA de la nostra base de dades, així com la variança explicada i altres conclusions.

Perquè ens va bé realitzar un PCA en el nostre problema de classificació?:

- Opció de reducció de variables (dimensionality curse).
- Interpretabilitat de les dimensions importants (associades a variables importants) i correlacions.
- Donar un primer cop d'ull al problema de separabilitat lineal de les classes.

Per a dur a terme l'anàlisi dels tres punts anteriors, mostrarem les figures de variància interpretada així com les projeccions per classe:

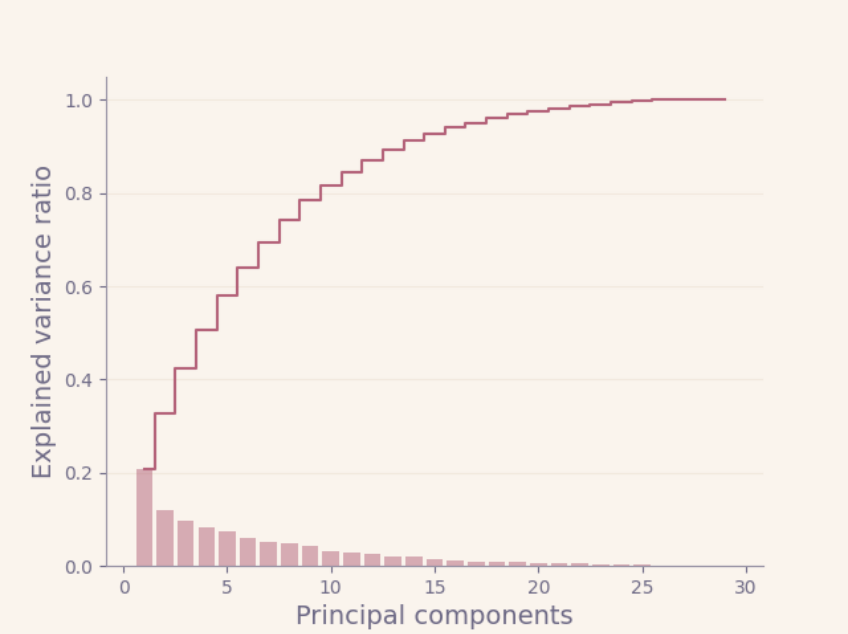


Figura 70: Variança explicada PCA

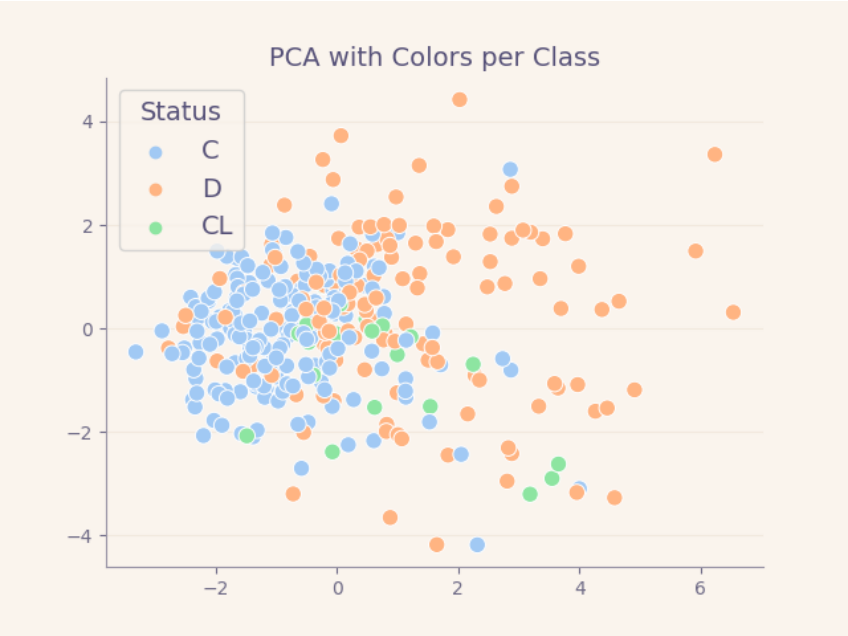


Figura 71: Projeccions per classe PCA

En la variància explicada podem observar com per arribar a una mesura raonable de 80% de variància explicada, necessitem unes 10 variables. Fet d'on podem extreure que no tenim un grup de variables molt selectes que mostrin la major part de la variància de la base de dades i les altres no tinguin importància. Sinó que moltes variables tenen el seu paper en explicar els diferents comportaments de les bases de dades.

Pel que fa a la projecció podem veure una dificultat de separació lineal especialment important entre les classes C i CL com anàvem comentant anteriorment. On tenim una concentració de les instàncies de la classe C que ens facilitarà la tasca.

Pel que fa a la reducció de variables i l'ús de les dimensions del PCA, el principal que ens tira enrere és la pèrdua d'interpretabilitat tan útil en l'àmbit mèdic. A més a més, veiem que ens haurem de quedar amb moltes dimensions per a poder representar correctament la base de dades (variança explicada), fet que ens tira encara més enrere.

L'observació de les variables més importants per dimensió ens mostra que la primera dimensió està majoritàriament representada per la Bilirubina mentre que la segona per les Plaquetes.

3 Definició de models

3.1 Experiments sobre les decisions presses

3.1.1 Escalat

En les seccions anteriors hem decidit utilitzar el StandardScaler per a escalar en els algorismes basats en distàncies (SVC i KNClassifier). Així doncs, anem a fer una petita prova en el train i sense balancejar per a veure quin mètode s'ajusta millor a les dades segons les mesures de F1 i balanced_accuracy. Els resultats han sigut els següents:

Classificador	F1	Balanced Accuracy	Accuracy
SVC sc	0.809348	0.595548	0.835329
SVC mmax	0.754940	0.543013	0.784431
KNN sc	0.753341	0.585097	0.763473
KNN mmax	0.684041	0.547025	0.703593

Taula 3: Comparació breu del mètode d'escalat

Amb aquest experiment simple, tot i que no podem afirmar res, ja que seria precipitat en experiment petit en el train, podem veure com a mínim que en el cas pitjor el mètode d'escalat no afecta en gran manera.

3.1.2 Outliers

En les seccions anteriors hem decidit utilitzar el StandardScaler per a escalar en els algorismes basats en distàncies (SVC i KNClassifier). Així doncs, anem a fer una petita prova en el train i sense balancejar per a veure quin mètode s'ajusta millor a les dades segons les mesures de F1 i balanced_accuracy. Els resultats han sigut els següents:

Classificador	F1	Balanced Accuracy	Accuracy
SVC sc	0.809348	0.595548	0.835329
SVC mmax	0.754940	0.543013	0.784431
KNN sc	0.753341	0.585097	0.763473
KNN mmax	0.684041	0.547025	0.703593

Taula 4: Comparació breu del mètode d'escalat

Amb aquest experiment simple, tot i que no podem afirmar res, ja que seria precipitat en experiment petit en el train, podem veure com a mínim que en el cas pitjor el mètode d'escalat no afecta en gran manera.

3.1.3 Outliers

En les seccions anteriors vam decidir no eliminar outliers per diverses raons. Ara bé, ara ho intentarem demostrar amb una petita prova amb el cross_val_score, tot i que serà sense classes balancejades, ens permetrà observar el comportament.

Com a metodologia, entrenarem dos algorismes amb assumpcions diferents com poden ser el SVC (potencial afectat pels outliers) i el DecisionTree (menys vulnerable a aquests) sobre un split amb outliers (els del IsolationTree, ja que és més fiable) i un sense. A continuació veurem la balanced accuracy i l'accuracy resultants fent la mitjana del 3-fold:

Classificador	Balanced Accuracy	Accuracy
SVC no_out	0.488140	0.724552
SVC out	0.501589	0.727317
KNN no_out	0.583873	0.721187
KNN out	0.581875	0.742466

Taula 5: Comparació eliminació outliers

Com podem observar, la mitjana dels 3 folds dona lleugerament millor lloc al dataset que conserva els outliers. Les raons pel qual això succeeix són explicades al detall en l'apartat d'Eliminació d'Outliers. Cal recalcar que és un experiment petit, s'espera que en un futur tingui més impacte.

3.1.4 Balanceig:

En aquest breu experiment observarem com afecta el tipus de balanceig als resultats, i si hem escollit bé el mètode SMOTE. Per això, provarem amb pesos, amb SMOTENC (per incloure-hi categòriques), o amb bootstrapping.

Mesurarem solament la mètrica de balanced_accuracy i accuracy en 10 folds i en el model SVC.

Classificador	Balanced Accuracy	Accuracy
SMOTENC	0.591550457339931	0.6970588235294118
weighting	0.6386189833558253	0.6943850267379681
Bootstrap	0.5904895786474735	0.6913547237076649

Taula 6: Comparació mètodes de balanceig

Veiem que era millor idea utiltizar weighting en els algorismes que ho permetin (tots menys KNClassifier). La diferència entre SMOTENC i Bootstrap no és significativament important.

3.2 Definició de metriques

Per a avaluar els diferents models, cal tenir en compte els diferents punts que podem avaluar en el nostre context en específic. En el nostre context ens caldria veure que tan bé prediu les classes en general, en segon lloc, com prediu les instàncies de classes minoritàries i, en tercer lloc, com es comporta el balanç precision/recall. Per això, s'ha escollit les següents mètriques:

- **Accuracy**: Mètrica escollida per a veure quantes instàncies prediu bé en general el nostre model.

- **Balanced Accuracy**: Mètrica que ens permetrà contrastar amb l'accuracy veient que tant prediu bé les instàncies de classes minoritàries, on descartarem models que es basin en prediure la classe majoritària.
- **F1 score**: Mètrica que ens permetrà veure el balanç precision/recall desitjat del nostre model.

Com podem veure, només amb aquestes tres mètriques podem fer-nos una idea de com funciona el nostre model en cada un dels aspectes a revisar.

3.3 Primer model triat: Decision Tree

El primer model triat és l'arbre de decisió. La motivació de la seva eleccio ve determinada per:

- **Interpretabilitat i explicabilitat:** Els arbres de decisió són per natura interpretables i fàcils d'entendre. Cada node de l'arbre representa una decisió basada en una característica, i cada branca representa els possibles resultats. Aquesta transparència és crucial en aplicacions mèdiques on entendre el procés de decisió és quasi tan important o més com les prediccions precises.
- **Complexitat:** Els arbres de decisió tenen una estructura senzilla i no solen fer overfitting quan es poda adequadament. Això és beneficiós quan es tracta d'un conjunt de dades de mida moderada com el nostre, el qual té 400 instàncies.
- **Hiperparàmetres**: Els arbres de decisió tenen hiperparàmetres que ens permeten controlar la complexitat de l'arbre, com la profunditat màxima de l'arbre o el nombre mínim de mostres necessàries per dividir un node. Això permet evitar l'overfiting/underfiting en les nostres dades.
- **Volum de dades:** Amb 400 instàncies, el conjunt de dades no és extremadament gran, on tot i que no és la situació idònia, els arbres poden tenir un bon paper.
- **Importància de variables:** Els arbres de decisió ens retornen puntuacions d'importància de les característiques, indicant quines d'aquestes són més influents en la presa de prediccions. Aquesta informació pot ser valuosa en un context mèdic, ja que ajuda a identificar els factors clau que contribueixen a quin serà l'estat del pacient.

3.3.1 Hiperparàmetres disponibles, utilitzats, i provats

Els hiperparàmetres que s'optimitzen generalment en el cas d'un DecisionTree són:

- **Criterion:** Funció per a mesurar la qualitat d'una divisió.
- **Max_depth:** Màxima profunditat de l'arbre de decisió.
- **Min_samples_split:** Mínimes instàncies requerides per a dividir un node intern.
- **Min_samples_leaf:** Mínimes instàncies requerides per a ser un node fulla.

Nosaltres utilitzarem totes les possibles combinacions de: $Criterion = ['giny','entropy']$, $max\_depth = ['None', 5, 10, 15]$, $min\_samples\_split = [2, 5, 10]$,$min\_samples\_leaf = [1, 2, 4]$. Ja que la taula està formada per 72 combinacions, no la mostrarem aquí, tot i que està present a la llibreta Jupyter.

Els primers hiperparàmetres per a realitzar el primer entrenament abans de fer el GridSearch seran els per defecte: criterion = gini, max_depth = None, min_samples_split = 2, min_samples_leaf = 1.

3.3.2 Primer entrenament(train)

Un primer entrenament abans de decidir els hiperparàmetres ens permetrà veure en què pot fallar el nostre model, en el nostre cas i segons les mètriques declarades, hem tingut 1 en accuracy, 1 en balanced_accuracy, i 1 en f1.

Podem veure a primera vista que el nostre model peca d'overfitting. Per a evitar això, haurem de realitzar una bona selecció d'hiperparàmetres en el següent apartat. No s'escau de corbes ROC ni matriu de confusió, ja que serien perfectes.

3.3.3 Anàlisi de resultats i iteració

Tot l'apartat d'anàlisi de resultats i iteració s'ha dut a terme mitjançant un GridSearchCV segons els hiperparàmetres comentats anteriorment. La preparació de variables ve automatitzada per una Pipeline i un ColumnTransformer. La mitjana dels resultats en els 10 folds corresponents han sigut els següents:

criterion	max_d	min_l	min_s	acc	b_acc	f1	std_acc	tr_acc
gini	NaN	1	2	0.679144	0.514264	0.514790	0.073497	1.000000
gini	NaN	2	2	0.676381	0.548343	0.545080	0.089210	0.944446
gini	15	1	2	0.676025	0.514549	0.511862	0.082081	0.999003
entropy	10	2	5	0.675847	0.588896	0.559194	0.073467	0.946105
entropy	NaN	2	2	0.675668	0.577126	0.545042	0.085436	0.958076
entropy	10	1	5	0.672995	0.543410	0.526048	0.075781	0.965740
entropy	15	2	5	0.672727	0.588567	0.557366	0.081538	0.950426
entropy	15	1	5	0.669875	0.541442	0.522794	0.085755	0.972722
entropy	NaN	1	5	0.669875	0.541442	0.522794	0.085755	0.973054
gini	10	1	2	0.667112	0.539627	0.532942	0.096259	0.973729
gini	10	2	2	0.667023	0.541550	0.537516	0.100635	0.932803
entropy	10	2	2	0.666667	0.568721	0.538107	0.082023	0.951757
entropy	NaN	2	5	0.666667	0.584151	0.552784	0.086138	0.950426
gini	15	2	2	0.664260	0.525409	0.525275	0.085512	0.943782
gini	15	1	5	0.664171	0.522747	0.503748	0.087565	0.964406
entropy	NaN	1	2	0.663904	0.537123	0.515918	0.078331	1.000000
entropy	10	1	2	0.663725	0.536216	0.517102	0.089689	0.988361
entropy	15	2	2	0.663547	0.568294	0.536019	0.091043	0.957744
entropy	15	1	2	0.660873	0.536100	0.514595	0.073940	1.000000
gini	NaN	1	5	0.658111	0.516907	0.498447	0.097146	0.965403
gini	5	1	2	0.654991	0.625641	0.570805	0.073779	0.794745
gini	15	2	5	0.654991	0.503498	0.499432	0.104300	0.933135

En la taula superior sols estan les 22 millors combinacions ordenades per accuracy. Tot i que tenim varietat, cal destacar el model amb giny, profunditat màxima 5, mínim suport per a fulla 1, mínim suport per a node 2. Aquest és el nombre 21 en accuracy, el nombre 4 en balanced_accuracy i el nombre 1 en f1 score. a més a més, té una desviació estàndard de 0.79(menor que la mitjana) i un 0.8 de accuracy en el train, fet que ens diu que no està fent overfit.

3.3.4 Resultats del primer model

El primer model escollit en els 10 folds realitzats, presenta un 66% d'accuracy i un 63% de balanced accuracy en els test. On solament té un 0.8 en els train i poca desviació estàndard. Fet que ens fa veure que no està fent gran overfiting.

Això ja dit, cal veure el seu comportament amb les classes minoritàries amb una matriu de confusió, així com les seves corbes AUC per a poder-lo validar. Ja que no tenim partició de val, s'ha fet sobre la de train.

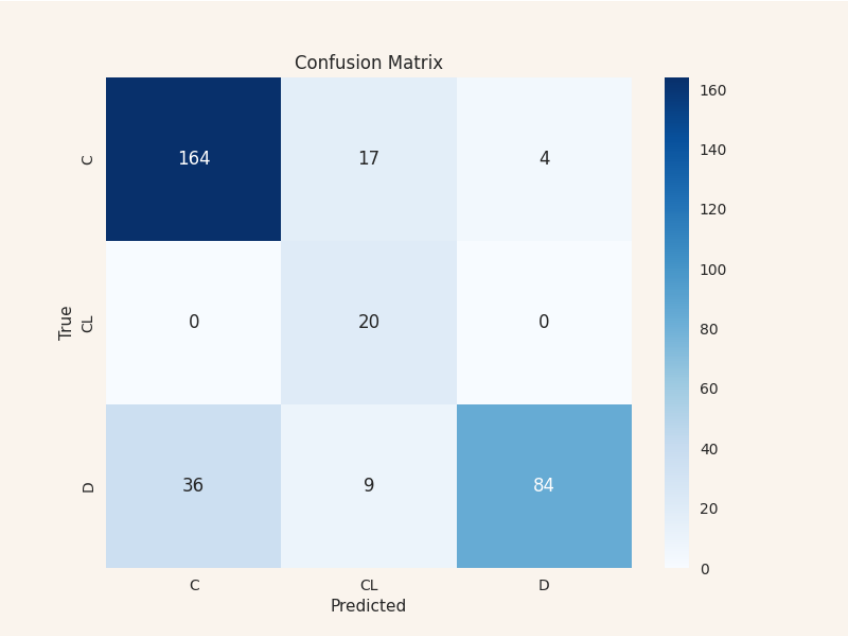


Figura 72: Matriu de confusió Decision Tree

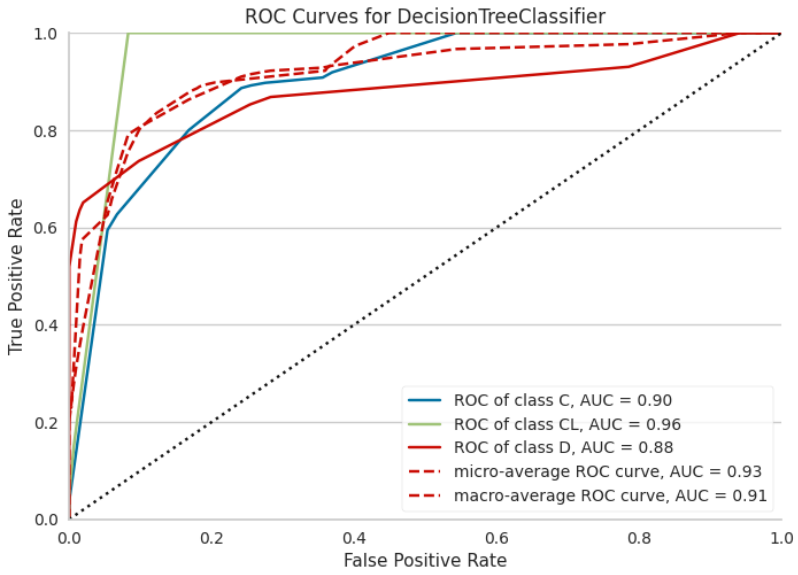


Figura 73: corbes ROC Decision Tree

Com podem veure, hem creat un predictor equilibrat que prediu de forma perfecta en el test la classe minoritària sense fer gaire overfit (vist al cross-validation).

Pel que fa a les corbes ROC, podem treure les mateixes conclusions, on els valors de l'AUC són favorables i tenim un molt bon rendiment quan el model és estricte.

3.4 Segon model triat: Support Vector Machine

El segon model triat és el SVM. La motivació de la seva elecció ve determinada per:

- **Interpretabilitat i explicabilitat:** Els SVM poden ser menys intuïtius i més difícils d'interpretar en comparació amb els arbres de decisió. La decisió es representa sovint en un espai d'alta dimensió, fent que sigui difícil de visualitzar directament.
- **Complexitat:** Els SVM són capaços de capturar relacions complexes en dades, incloent-hi patrons no lineals, utilitzant diferents funcions com la de base radial (RBF). La qual usarem com veurem més endavant.
- **Hiperparàmetres**: Els SVMs tenen hiperparàmetres com l'elecció del kernel, el paràmetre de regularització (C), i, fins i tot paràmetres corresponents a la funció del kernel (com pot ser la gamma). Aquests ens permetràn controlar l'overfitting (especialment important en la grandària de la nostra base de dades).
- **Volum de dades:** Els SVMs poden funcionar en conjunts de dades de mida petita o mitjana de manera efectiva. En el cas de 400 casos com tenim, el SVM hauria d'anar força bé, especialment si el nombre de característiques no és excessivament alt (on en el nostre cas en són 30). Per a conjunts de dades més grans, l'entrenament d'un SVM pot ser molt costós, però amb la mida del conjunt de dades proporcionada, no hauria de ser un problema.
- **Convergència:** La convergència pot ser un problema, especialment quan no hem eliminat els outliers presents a la nostra base de dades.

3.4.1 Hiperparàmetres disponibles, utilitzats, i provats

Els hiperparàmetres que s'optimitzen generalment en el cas d'un SVM són:

- **C:** Coeficient the regularització
- **kernel:** Tipus de kernel utilitzat en l'algorisme
- **gamma:** Coefficient de kernel per a 'rbf','poly' i 'sigmoid'

Nosaltres utilitzarem totes les possibles combinacions de: $C = [0.1, 1, 10, 100]$, $kernel = ['linear','rbf']$, i $\gamma = ['scale','auto',0.1,0.01]$. On totes les combinacions provades són les de la taula següent:

Taula 7: Combinacions grid_search

#	param_svc__C	param_svc__gamma	param_svc__kernel
11	1	auto	rbf
0	0.1	scale	linear
2	0.1	auto	linear
4	0.1	0.1	linear
15	1	0.01	rbf
6	0.1	0.01	linear
23	10	0.01	rbf
3	0.1	auto	rbf
1	0.1	scale	rbf
12	1	0.1	linear
10	1	auto	linear
9	1	scale	rbf
8	1	scale	linear
14	1	0.01	linear
13	1	0.1	rbf
5	0.1	0.1	rbf
19	10	auto	rbf
7	0.1	0.01	rbf
28	100	0.1	linear
18	10	auto	linear
26	100	auto	linear
17	10	scale	rbf
30	100	0.01	linear
22	10	0.01	linear
16	10	scale	linear
31	100	0.01	rbf
20	10	0.1	linear
24	100	scale	linear
21	10	0.1	rbf
29	100	0.1	rbf
27	100	auto	rbf
25	100	scale	rbf

Els primers hiperparàmetres per a realitzar el primer entrenament abans de fer el GridSearch seran els per defecte: C = 1, kernel = 'rbf', gamma = 'scale'.

3.4.2 Primer entrenament(train)

Un primer entrenament abans de decidir els hiperparàmetres ens permetrà veure en què pot fallar el nostre model, en el nostre cas i segons les mètriques declarades, hem tingut 0.82 en accuracy, 0.58 en balanced_accuracy, i 0.59 en f1.

Podem veure a primera vista un bias cap a la classe majoritària en el balanced_accuracy, així com un balanç regular del precision/recall.

Veurem la matriu de confusió i corba ROC per comparar amb els hiperparàmetres escollits posteriorment:

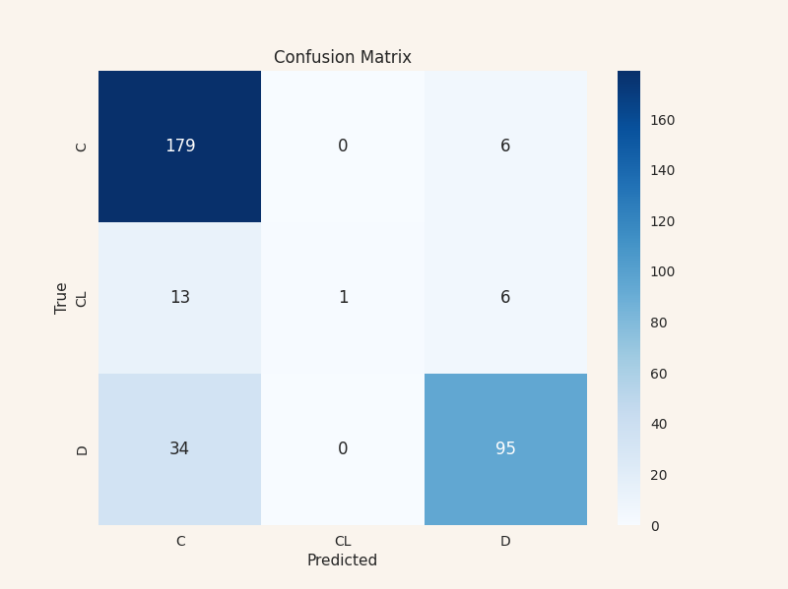


Figura 74: Matriu de confusió SVM inicial

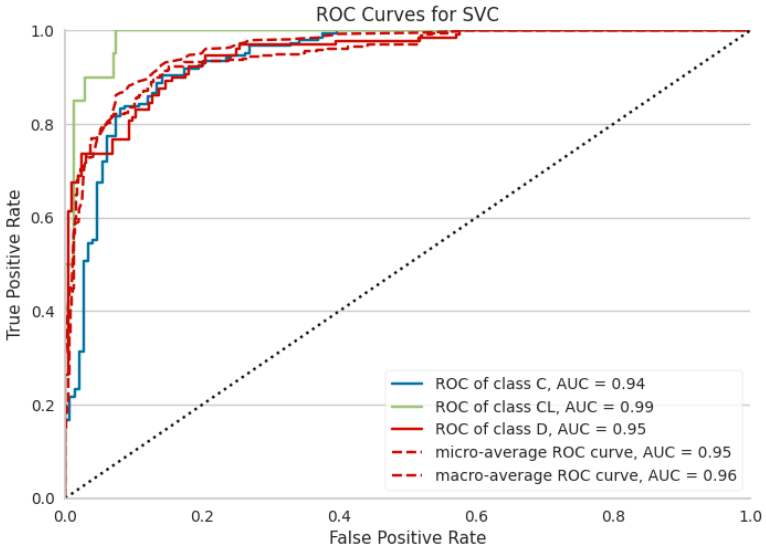


Figura 75: corbes ROC SVM inicial

Com hem dit, veiem a la matriu de confusió com quasi mai hem predit la classe minoritària, tenim un bias cap a la classe majoritària en tots els casos. Tot i això, les corbes ROC surten prou bé, segurament es pel fet que ens trobem en la partició de train.

3.4.3 Anàlisi de resultats i iteració

Tot l'apartat d'anàlisi de resultats i iteració s'ha dut a terme mitjançant un GridSearchCV segons els hiperparàmetres comentats anteriorment. La preparació de variables ve automatitzada per una Pipeline i un ColumnTransformer. La mitjana dels resultats en els 10 folds corresponents han sigut els següents:

C	gamma	kernel	acc	b_acc	f1	std_acc	tr_acc
1	0.100000	rbf	0.700178	0.598182	0.583405	0.070999	0.898528
100	0.010000	rbf	0.694207	0.594156	0.561116	0.064441	0.940110
10	0.100000	linear	0.691444	0.590276	0.565401	0.077558	0.811370
10	auto	rbf	0.687968	0.600986	0.572452	0.078429	0.941105
100	0.100000	linear	0.685383	0.586767	0.558534	0.081720	0.818014
10	0.010000	rbf	0.685294	0.642660	0.580650	0.058549	0.803382
100	auto	rbf	0.685205	0.567964	0.551328	0.073757	0.999334
10	scale	linear	0.685205	0.570138	0.547122	0.080138	0.806374
1	scale	rbf	0.682175	0.627320	0.591126	0.057913	0.831326
10	scale	rbf	0.682086	0.552009	0.535407	0.103734	0.984699
100	0.010000	linear	0.679323	0.582351	0.550544	0.078318	0.823338
100	scale	linear	0.679234	0.593908	0.570415	0.078801	0.823330
100	auto	linear	0.676381	0.579982	0.554171	0.080340	0.823007
10	0.010000	linear	0.676203	0.577242	0.551205	0.077260	0.813029
10	0.100000	rbf	0.673351	0.501316	0.497505	0.096893	0.999334
100	0.100000	rbf	0.673173	0.500409	0.496934	0.094911	1.000000
10	auto	linear	0.670053	0.559241	0.532810	0.087068	0.811032
0.100000	0.010000	linear	0.669964	0.660699	0.580801	0.068909	0.737514
0.100000	0.100000	linear	0.669964	0.660699	0.581681	0.074048	0.737846
1	0.010000	linear	0.667112	0.602418	0.553027	0.055714	0.780092
1	0.100000	linear	0.664082	0.600566	0.547897	0.061159	0.780424
0.100000	scale	linear	0.663904	0.655473	0.576959	0.077563	0.738846

En la taula superior sols estàn les 22 millors combinacions rankejades per accuracy. Tot i que tenim varietat, cal destacar el model C=10, gamma = 0.01 i kernel = 'rbf'. És el sisè millor en les tres mètriques. A més a més, presenta una de les desviacione estàndard més baixes en la accuracy i un accuracy solament del 0.8 en el train, fet que ens indica que no està fent molt overfit.

3.4.4 Resultats del segon model

El segon model escollit en els 10 folds realitzats, presenta un 69% d'accuracy i un 64% de balanced accuracy en els test. On solament té un 0.8 en els train i poca desviació estàndard. Fet que ens fa veure que no està fent gran overfiting.

Això ja dit, cal veure el seu comportament amb les classes minoritàries amb una matriu de confusió, així com les seves corbes AUC per a poder-lo validar. Ja que no tenim partició de val, s'ha fet sobre la de train.

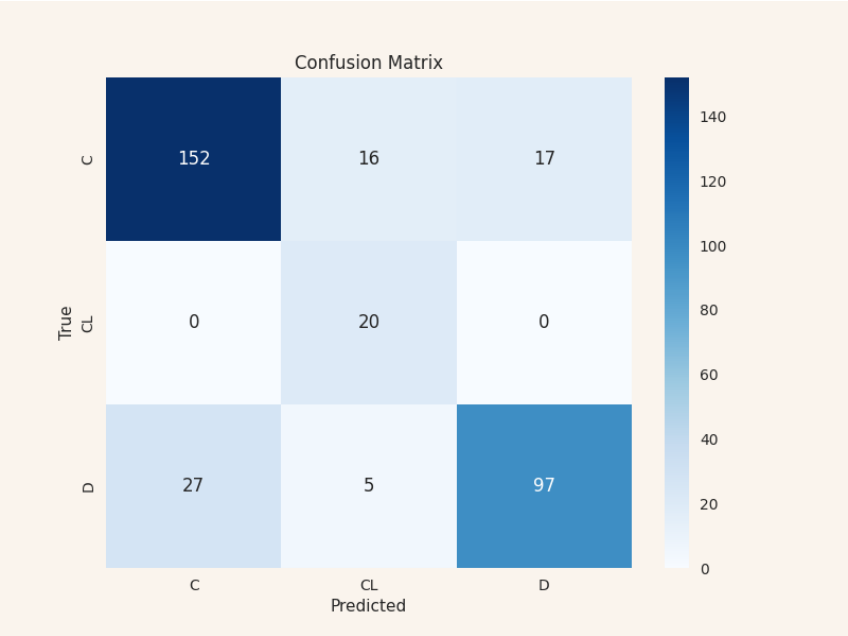


Figura 76: Matriu de confusió SVM final

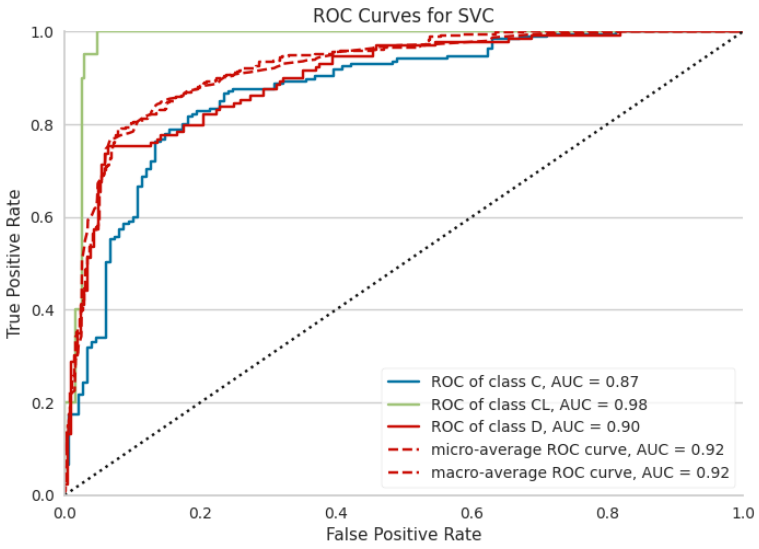


Figura 77: corbes ROC SVM final

Com podem veure, hem creat un predictor equilibrat que una vegada més prediu de forma perfecta la classe minoritària sense fer gaire overfit (vist al cross-validation). Ara bé, fallem més en prediure la C i la D

Pel que fa a les corbes ROC, podem treure les mateixes conclusions, on els valors de l'AUC són favorables i tenim un molt bon rendiment quan el model és estricte.

Podem veure una millora radical amb el tractament de la classe minoritària tan important en el sector mèdic respecte abans de tractar els hiperparàmetres.

3.5 Tercer model triat: KNN

El tercer model triat és el KNN. La motivació de la seva eleccio ve determinada per:

- **Interpretabilitat i explicabilitat:** Explicar prediccions pot ser un repte quan tenim moltes dimensions, i les característiques específiques que contribueixen a una predicció poden no ser fàcilment aparents.
- **Complexitat:** KNN és un model no paramètric i basat en instàncies. No fa moltes suposicions sobre la distribució de dades. És relativament simple, fet que ens va bé.
- **Hiperparàmetres**: El principal hiperparàmetre és k (nombre de veïns) on també tenim 'weights', que prioritza les mostres més properes, i 'p' que ens diu la potència per la mètrica de Minkowski. Ens va bé que quasi tota la importància dels hiperparàmetres en recaigui sols sobre un.
- **Volum de dades:** Amb un conjunt de dades de 400 instàncies, KNN hauria de ser factible. Si en tinguessim més podria ser un problema pel seu cost.
- **Imbalanç:** KNN és especialment sensible a desequilibri de classes, això és degut en gran part perquè si no hi ha instàncies de la teva classe, no les trobaràs com a veïnes. Ja que hem balancejat les nostres dades, no esdevé un gran problema.

3.5.1 Hiperparàmetres disponibles, utilitzats, i provats

Els hiperparàmetres que s'optimitzen generalment en el cas d'un SVM són:

- **n_neighbours:** k, nombre de veïns.
- **p:** potència de la mètrica de Minkowski
- **weights:** Pes donat a les diferents instàncies.

Nosaltres utilitzarem totes les possibles combinacions de: $n\_neighbours = [3, 5, 7]$, $p = [1, 2]$, i $weights = ['uniform', 'distance']$. On totes les combinacions provades són les de la taula següent:

n_neighbors	p	weights
7	2	distance
7	2	uniform
5	2	distance
7	1	uniform
7	1	distance
5	2	uniform
5	1	uniform
5	1	distance
3	1	distance
3	2	distance
3	1	uniform
3	2	uniform

3.5.2 Primer entrenament(train)

Un primer entrenament abans de decidir els hiperparàmetres ens permetrà veure en què pot fallar el nostre model, en el nostre cas i segons les mètriques declarades, hem tingut 0.8 en accuracy, 0.85 en balanced_accuracy, i 0.74 en f1. Resultats prou bons

Podem veure a primera vista un molt bon comportament d'aquest model simple, tot i això haurem de realitzar una bona selecció d'hiperparàmetres en el següent apartat per a millorar els resultats.

Veurem la matriu de confusió i corva ROC per comparar amb els hiperparàmetres escollits posteriorment:

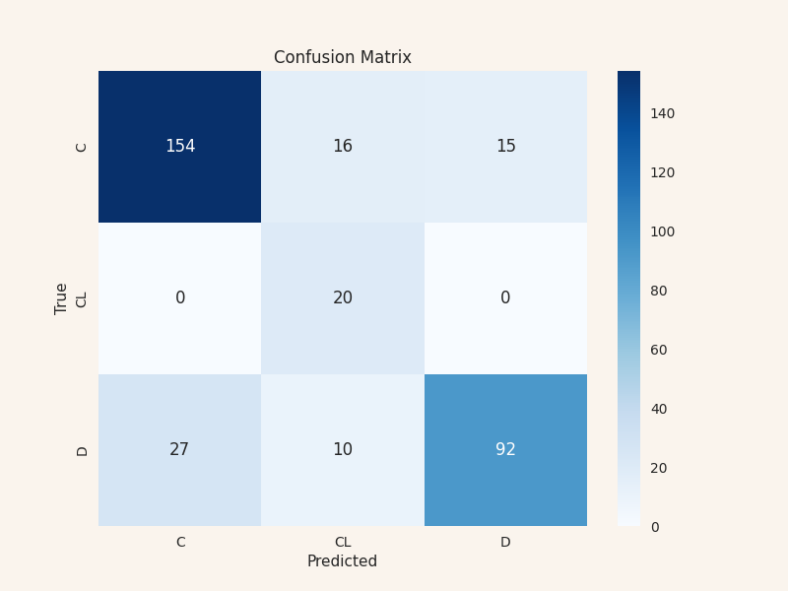


Figura 78: Matriu de confusió KNN inicial

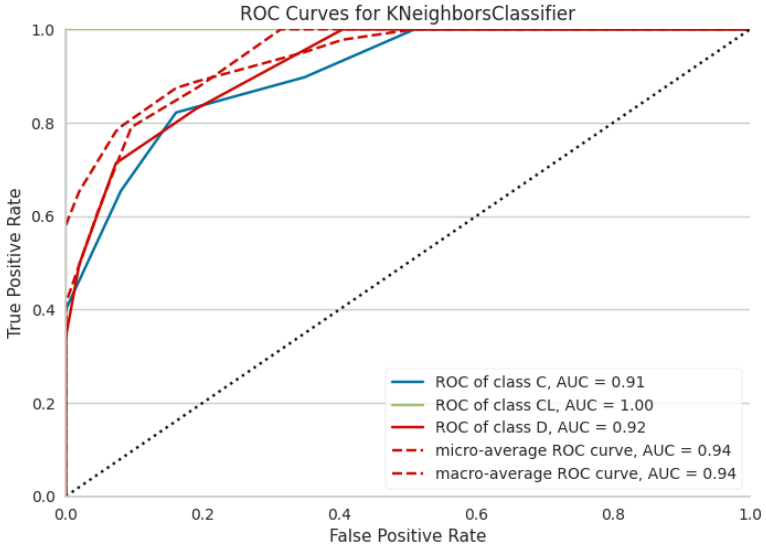


Figura 79: corbes ROC KNN inicial

Com hem dit, veiem a la matriu de confusió com quasi mai predim la classe minoritària, tenim un bias cap a la classe majoritària en tots els casos. Tot i això, les corbes ROC surten prou bé, segurament és pel fet que ens trobem en la partició de train.

3.5.3 Anàlisi de resultats i iteració

Tot l'apartat d'anàlisi de resultats i iteració s'ha dut a terme mitjançant un GridSearchCV segons els hiperparàmetres comentats anteriorment. La preparació de variables ve automatitzada per una Pipeline i un ColumnTransformer. La mitjana dels 10 resultats en els folds corresponents han sigut els seguents:

n_neighbors	p	weights	accuracy	balanced_accuracy	f1	std_accuracy
7	2	distance	0.631569	0.598183	0.654682	0.059037
7	2	uniform	0.628539	0.597143	0.653592	0.066963
5	2	distance	0.652601	0.596898	0.670465	0.055957
7	1	uniform	0.637494	0.583770	0.658347	0.061601
7	1	distance	0.631434	0.581031	0.653835	0.070447
5	2	uniform	0.649661	0.579367	0.668938	0.064626
5	1	uniform	0.631615	0.578085	0.645041	0.069141
5	1	distance	0.646540	0.573856	0.656322	0.074362
3	1	distance	0.646585	0.548040	0.652327	0.045865
3	2	distance	0.646676	0.546413	0.648833	0.058035
3	1	uniform	0.640751	0.539863	0.641063	0.064508
3	2	uniform	0.634781	0.522157	0.639980	0.058313

Quant als resultats d'aquest cross-validation podem veure una menor significança dels hiperparàmetres a l'hora de predir els resultats. Tot i això, podem veure com destaca en les diferents mètriques la combinació d'hiperparàmetres n_neighbours = 5, p = 2, weights = 'distance'. Amb la millor accuracy, tercera millor balanced_accuracy (per molt poc), i primer lloc en f1. Pel que fa al seu fit, podem veure que la seva desviació mitjana en els diferents folds és la segona més baixa, indicant relativament poca variància entre prediccions en els folds. Per tant, ja tenim combinació per al KNN.

3.5.4 Resultats del tercer model

El segon model escollit en els 10 folds realitzats, presenta un 65% d'accuracy , un 60% de balanced accuracy i un 67% de f1 en els tests. On solament té una desviació estàndard de 0.055. Fet que ens fa veure que no està fent gran overfiting.

El model, però amb la nova mètrica té una performance perfecte al train. Així doncs, no escau fer ni matriu ni confusió ni corba ROC

Fora de ser un mal indicador, cal recordar que estem actuant al split de train i no podem tenir gaires conclusions. És possible que funcioni força bé.

4 Selecció de model

4.1 Descripció del model triat.

Després de veure el comportament dels diversos models, he decidit escollir-ne dos: L'arbre de decisió i el SVM.

Per què? Mentre que el SVM ens dona millors resultats i tendeix a generalitzar millor, l'arbre de decisió serà molt més útil a l'hora de poder treballar amb personal mèdic per la seva explicabilitat. Que, com hem recalcat, és essencial en aquest àmbit. Els motius individuals de cada mètode ja han sigut discutits en l'apartat 3.

Segons el procés iteratiu fet anteriorment, els hiperparàmetres escollits per al SVM són C= 10, gamma = 0.01 i rbf = kernel. Mentre que per al DecisionTree hem escollit criterion = 'gini', max_depth = 5, min_samples_split = 2, min_samples_leaf = 1.

4.2 Anàlisi de les limitacions i capacitats del model

Per l'origen i natura de les nostres dades, el nostre dos models estaran limitats per:

- **Pacients amb cirrosi**
- **Principalment femelles**
- **Adults**
- **Principalment gent d'Estats Units**
- **Desfassament temporal:** Estudi fet a mitjans dels anys 80.

Quant a les capacitats:

- **Molta dificultat en predir la gent que tindrà trasplantament:** És una categoria poc diferenciable de la supervivència, on a més teníem molt poques mostres.
- **Dificultat relativa en predir la gent que morirà.**
- **Capacitat de predicció moderada:** Ens trobem pel voltant del 0.7 d'accuracy i balanced_accuracy. No és excessivament fiable.
- **Desfassament temporal:** Com ja hem dit, les capacitats predictives es veuran molt reduïdes per aquest fenòmen

4.3 Resultats en partició de test, en comparació amb train i val

Fins a aquest punt, no hem tocat en cap moment les dades tel test. Així doncs, ara cal comparar com es comporta el model en veure dades fora de les que teníem. Les taules següents marquen el rendiment en les tres mètriques dels dos models.

SVM:

Taula 8: Resultats SVM diferents splits

Dataset	Accuracy	Balanced Accuracy	F1 Score
Train	0.81	0.86	0.77
Validation	0.69	0.64	0.58
Test	0.77	0.54	0.54

Decision Tree:

Taula 9: Resultats DecisionTree diferents splits

Dataset	Accuracy	Balanced Accuracy	F1 Score
Train	0.8	0.83	0.74
Validation	0.65	0.63	0.57
Test	0.67	0.52	0.52

Podem observar una lleugera baixada en el rendiment en les particions de test. Especialment, veiem com, en tenir molt poques mostres de la classe CL, els del test segurament tenien poc a veure amb els del train. Fet que ha provocat una baixada considerable en la balanced accuracy i f1 en els models tot i que les estàvem prioritzant a tot cost.

Mirarem també les matrius de confusió i corbes ROC del model per a confirmar les nostres hipòtesis:

4.3.1 Descriptiva prediccions SVM.

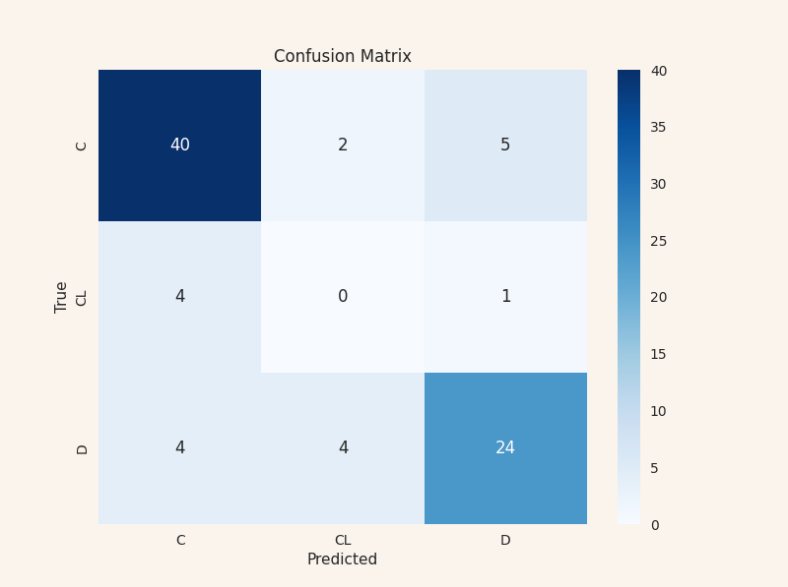


Figura 80: Matrius de confusió test SVM

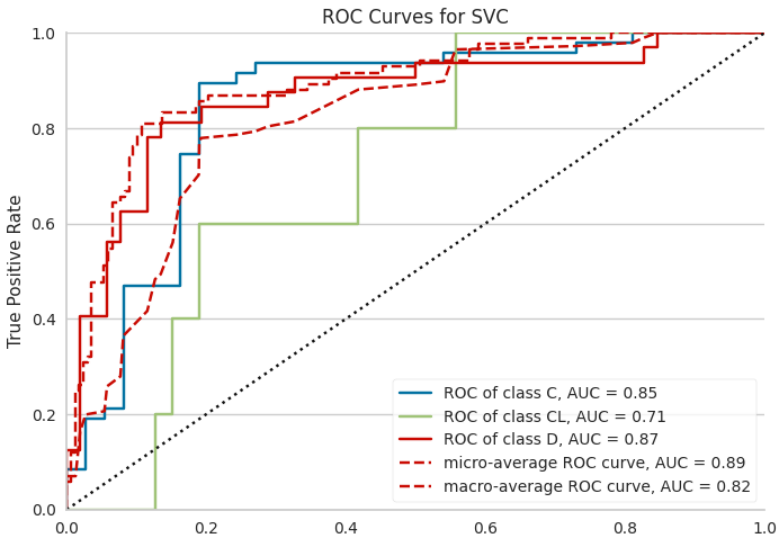


Figura 81: corbes ROC test SVMl

Veiem que no hem predit bé ni una sola persona amb trasplantament, on les altres dues classes s'han predit força bé com podem veure als ROCS. El reduït format del test, però, pot haver influenciat les mètriques i aquest estudi contundentment, així com la avaluació de generalització dels models.

4.3.2 Descriptiva prediccions DecisionTree.

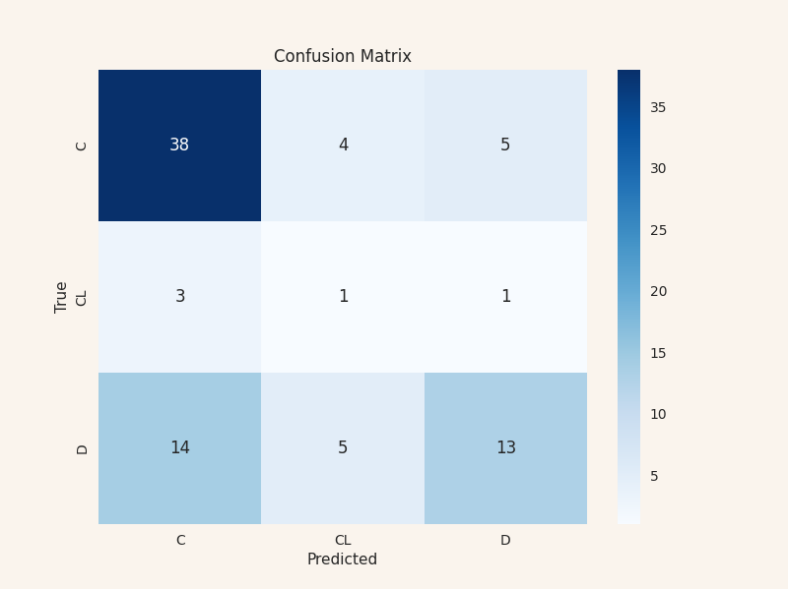


Figura 82: Matrius de confusió test DecisionTree

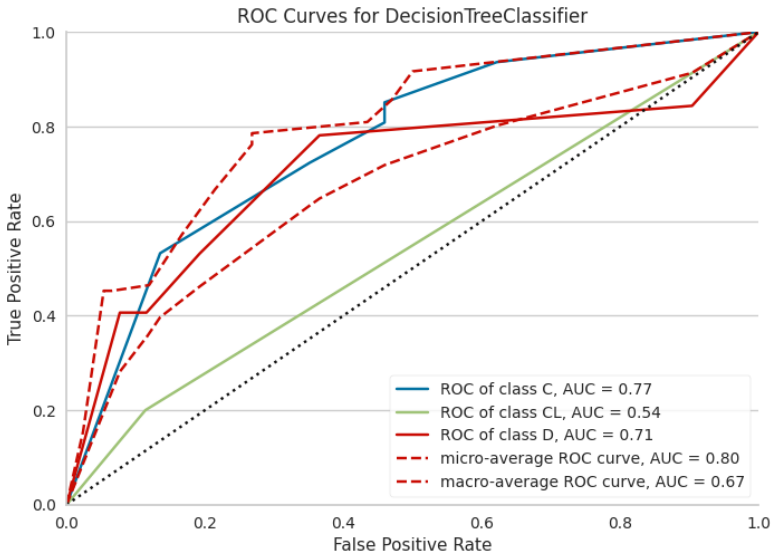


Figura 83: corbes ROC test DecisionTreel

En aquest model, però, podem veure un rendiment molt pobre en totes les classes, on segurament no hem sabut capturar les complexitats o un arbre per si sol no ha pogut. Problema que no ha tingut la SVM. Tot i els seus mals resultats, potser seria útil analitzar les seves separacions i examinar tot el que ens pot donar la seva interpretabilitat.

5 Bonus-EBM

En aquest apartat entrenarem una EBM, comparant els seus resultats amb la dels altres models. Així com veient la importància de les variables i com les ha tractat.

5.1 Resultats

Quant als resultats, podem veure a la taula de baix com ha tingut uns grans resultats en el test comparat amb els altres dos models que hem provat en aquella partició. Particularment ha millorat molt el precision/recall reflectat en el f1, tot i que no ha millorat molt la balanced_accuracy (segurament per la falta de mostres):

Taula 10: Taula rendiment test

Model	Performance Metrics		
	Accuracy	Balanced Accuracy	F1
SVM	0.77	0.54	0.54
DecisionTree	0.67	0.52	0.52
EBM	0.74	0.53	0.74

Pel que fa a l'AUC i les corbes roc, podem observar que la classe CL surt perdent una altra vegada, on els altres dos AUC funcionen relativament bé. Podem veure que té un rendiment similar a la SVM però amb el gran bonus de l'explicabilitat.

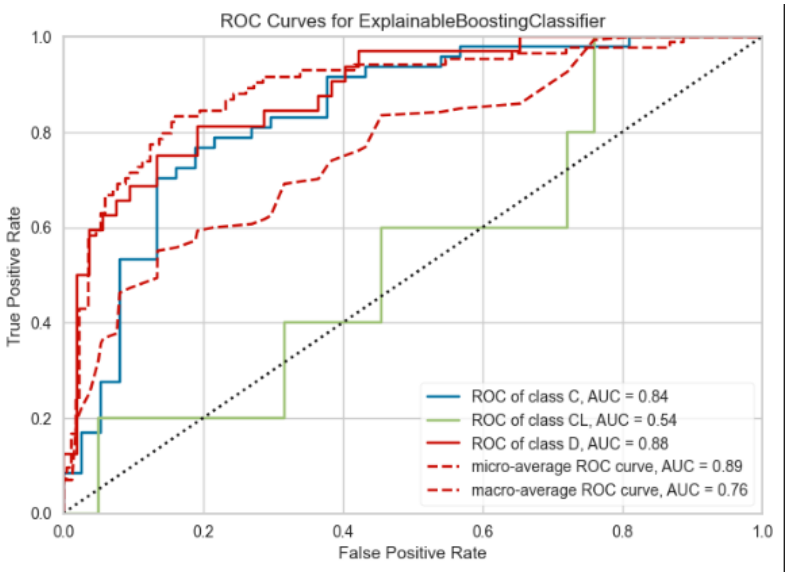


Figura 84: corbes ROC test EBM

Veiem a la matriu de confusió com no prediu bé cap instància de la classe CL, mentre les altres dos les prediu força bé.

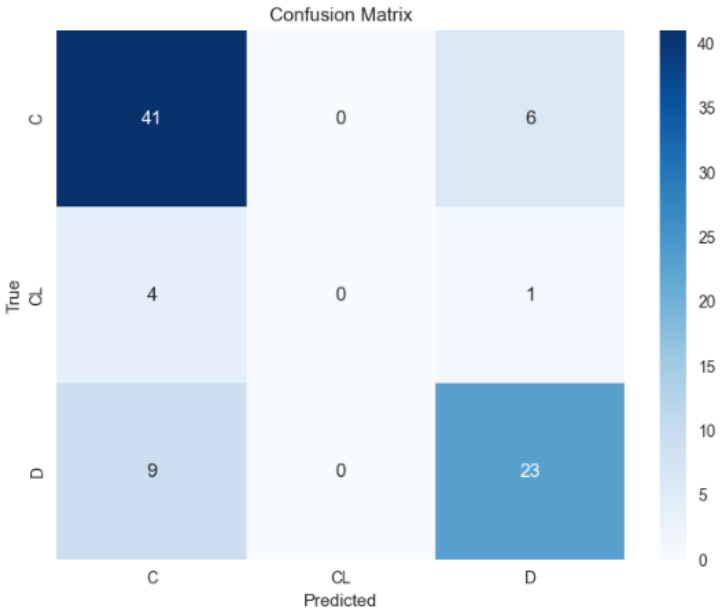


Figura 85: matriu de confusió test EBM

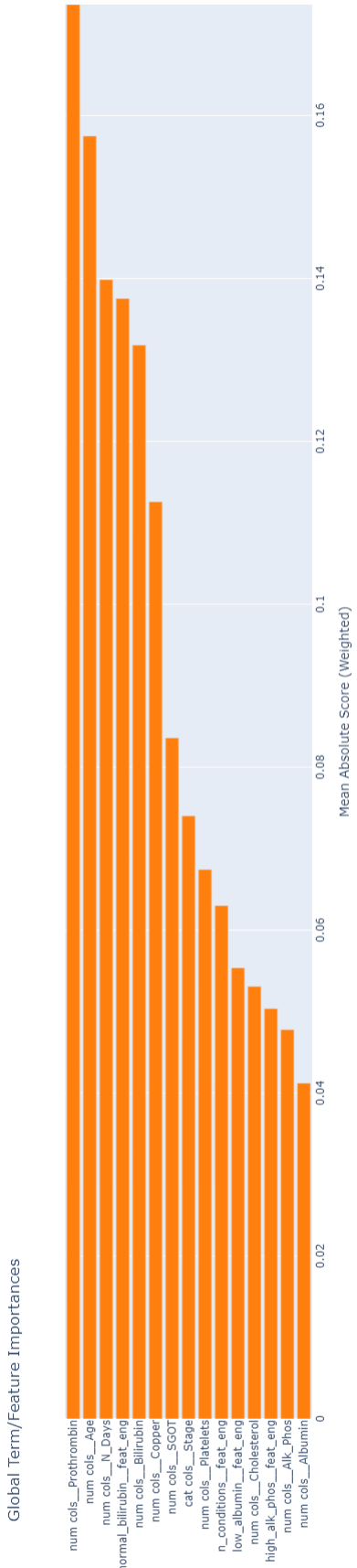
5.2 Importància variables:

Veiem en el gràfic inferior com les variables Prothrombin, Age,N_days i Bilirrubin són les més informatives. També podem veure com les variables creades a través de l'enginyeria de variables són bastant informatives.

Taula 11: Rànquing variables EBM

Rank	Variable
1	Prothrombin
2	Age
3	N_days
4	normal_bilirrubin
5	Bilirrubin
6	Copper
7	SGOT
8	Stage
9	Platelets
10	n_conditions
11	low_albumin
12	Cholesterol
13	high_alk_phos
14	Alk_Phos
15	Albumin

Aquests resultats, fora de ser estrany, concorden clarament amb el que hem vist fins ara, mitjanes per classe, distribucions, etc. També podem concloure una vegada més que hi ha moltes variables influents, on no tot recau sobre una o dues.


Global Term/Feature Importances
num cols__Prothrombin
num cols__Age
num cols__N_Days
normal_bilirubin__feat_eng
num cols__Bilirubin
num cols__Copper
num cols__SGOT
cat cols__Stage
num cols__Platelets
n_conditions__feat_eng
low_albumin__feat_eng
num cols__Cholesterol
high_alk_phos__feat_eng
num cols__Alk_Phos
num cols__Albumin
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Mean Absolute Score (Weighted)

6 Bonus- Clustering:

Realitzar clústering sobre les nostres dades ens pot permetre identificar colûlectius crucials que ens poden servir en la nostra tasca i fora d'aquesta.

Per acomplir aquesta tasca, hem realitzat dos mètodes diferents amb també diferent nombre de clusters per a veure diferents comportaments. Els mètodes són KMeans i Hierarchical Clustering

6.1 KMeans

Pel que fa al KMeans, hem volgut posar k = 3 per a veure si podíem trobar les tres classes que teníem a prediure. Després d'aplicar l'algorisme, teníem aquestes tres classes:

Figura 86: PCA kmeans

Podem veure que sobretot la distinció de la classe 0 amb la classe 1 i 2 és molt similar a la de la classe D amb la C i CL. Mentre que la distinció entre C i CL es fa molt més difícil.

Respecte al profiling de les tres classes realitzat a partir de boxplots per classe (present a la llibreta Jupyter) podem descriure-les com:

- **Classe 0:** Gent greu, amb les variables Prothtombina, Bilirrubin i Cholesterol alts, variables que hem vist que són molt influents a l'EBM. Comparable amb la classe D de la classificació.
- **Classe 1:** Gent amb condicions derivades per la cirrosi, el més probable és que estan controlats, ja que tenen els indicadors bastant estables.
- **Classe 2:** Punt intermedi entre la classe 0 i 1, fases intermèdies, indicadors estables però una mica pitors que la classe 1. Possiblement, s'haurien de tractar

D'aquí en podem treure informació important, pel fet que identificar en quina comunitat està cada persona ens pot indicar en quin punt de la afecció està cadascú, o potser en el cas de la gent de la classe 2 si les hem de tractar. Podem veure-ho d'una manera visual, on si tiréssim les instàncies verdes cap a l'esquerra (baixar indicadors que representen les dimensions com Bilirrubina) podriem separar gent molt greu de gent controlada.

6.2 Jeràrquic

En quant al KMeans, hem fet el dendograma i tallat en k = 5, ja que presentava el millor equilibri en distància(eix y) i nombre de clústers com es pot veure a la figura:

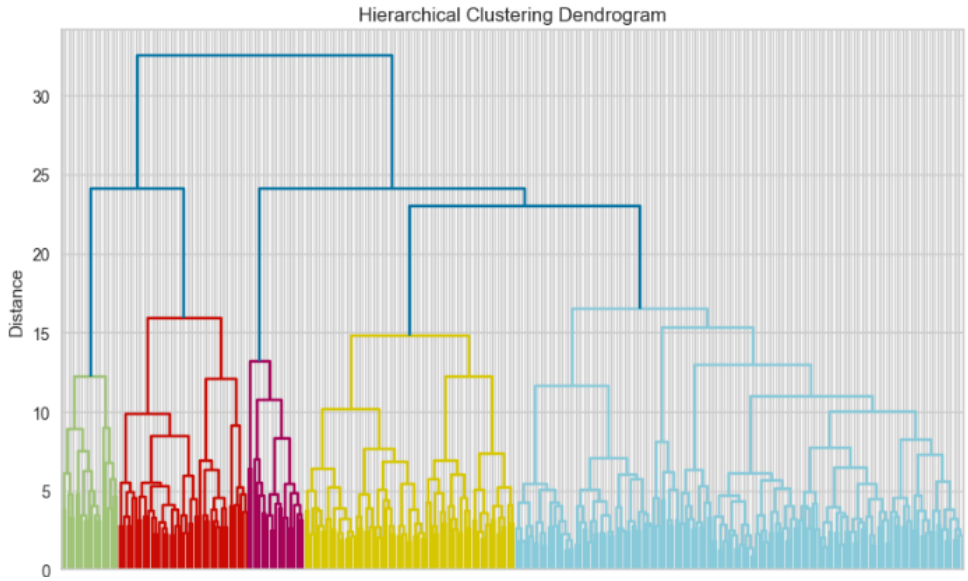


Figura 87: Dendograma amb tall k= 5

Podem veure com tenim una classe majoritària, dues mitjanes i dues minoritàries. Per a veure que interpreta cada una, una altra vegada més hem realitzat boxplots per classe. Podem descriureles com:

- **Classe 0:** Gent greu, etapes finals i grans nivells de bilirubina i colesterol.
- **Classe 1:** Gent gran amb els indicadors malament, tot i no tenir moltes malalties derivades de la cirrosi, fosfat alcalí pels núvols. N_days ens fa entendre que o bé els fan trasplantaments ràpids o bé moren.
- **Classe 2:** Gent amb N_days molt elevat, o no han mort fins a la data d'estudi, o els hi han fet un trasplantament. És la classe amb els millors indicadors, tot i que estan a fases altes. Fa pensar que estan controlats.
- **Classe 3:** Gent a les últimes fases i amb moltes condicions, valors no molt bons en quasi tots els indicadors.
- **Classe 4:** Grup estable amb valors mitjans entre tots els indicadors, i majoritàriament dins dels rangs saludables.

En aquest clustering se'ns fa molt difícil caracteritzar als individus, una utilitat potser recauria en si caracteritzem un individu de la classe 3 controlar-lo per passar a la classe 2, fent baixar els indicadors perquè no desenvolupin tantes condicions.

Com hem vist, trobem utilitats fora del treball, podent caracteritzar a pacients per a tractar-los adequadament. També, però, podríem haver utilitzat aquesta informació a l'hora de generar els models, donant informació addicional dels grups als quals pertanyen segons l'anàlisi no supervisat (model stacking). Fet que segurament milloraria el rendiment i les mètriques.

7 Model Card for Cirrhosis Patient Survival Prediction

Model Details

Overview

Aquest model prediu si un cert pacient amb cirrosi mor, necessita trasplantament de fetge, o sobreviu sense trasplantament. La base de dades prové de Fleming, Thomas R., and David P. Harrington. Counting processes and survival analysis. Vol. 625. John Wiley & Sons, 2013, està formada per 418 pacients de cirrosi. Està entrenat amb un algorisme d'EBM el qual és un Model Additiu generalitzat amb detecció automàtica d'interaccions. Els hiperparàmetres que s'ha hagut de controlar és el nombre dinteractions = 5.

Version

name: 4bd78be7-3c1e-4e30-ac3f-2bb0a9bc8d5c
date: 2023-12-28

Owners

Jordi Granja Bayot, jordi.granja.bayot@gmail.com

References

- https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+
- https://www.semanticscholar.org/paper/Prognosis-in-primary-biliary-cirrhosis%3A-Model-for-Dickson-Grambsch/db1487216b8a4b26f5e5078ea9109ac9d8355b65
- https://scikit-learn.org/stable/

Considerations

Intended Users

- Estudiants de IAA
- Professors de IAA

Use Cases

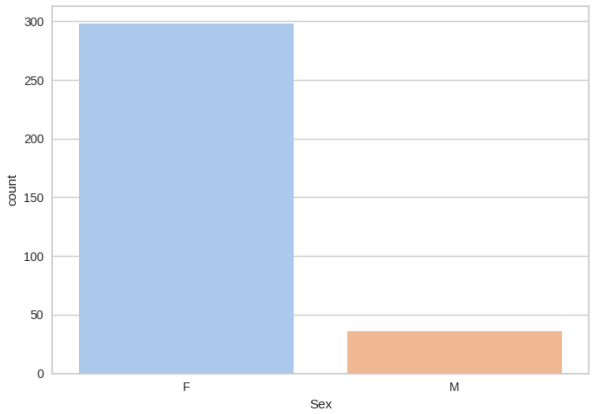
- L'intenció d'ús d'aquest model és el d'agafar experiència i ser avaluat mitjançant una tasca d'aprenentatge supervisat en el qual es pretén classificar pacients que pateixen de cirrosi en els seus respectius estats. Aquest model no té cap intenció en ser usat per a realitzar cap diagnòstic real o negoci. Els usuaris que poden fer servir aquest model es troben sols entre el professorat i alumnat.

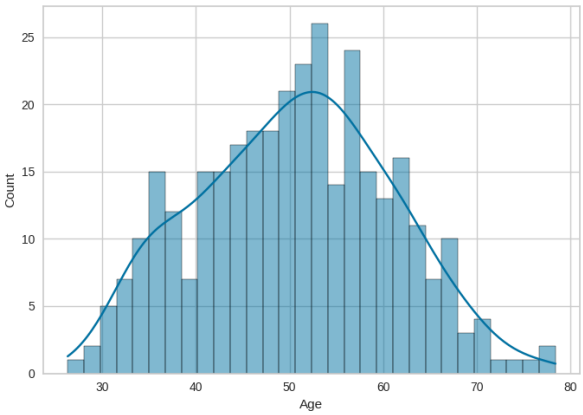
Limitations

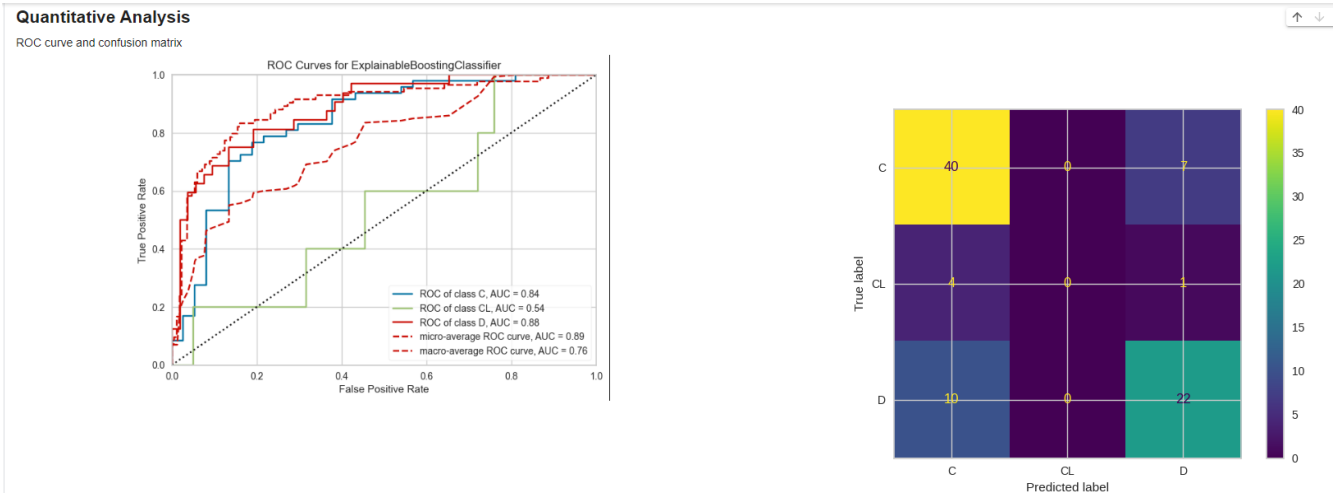
- El model està limitat principalment a pacients amb cirrosi, principalment femelles, adults, i gent d'Estats Units. També està limitat per l'època en la qual es va fer l'estudi (anys 80). El model no funciona bé en gent que necessiti trasplantament o que estigui relativament greu. Tampoc es garantitza la seva plena utilitat en gent no adulta, homes, gent d'avui dia (desfassament temporal). El pas del temps pot degradar encara més el possible ús daquest model. Així com lús de diferents eines per a mesurar els diferents indicadors.

Ethical Considerations

- Risk: La selecció manual de pacients pot crear un selection bias
  Mitigation Strategy: Estratificar la selecció de pacients dins del possible segons la població general
- Risk: Si mai es fan prediccions en làmbit mèdic (fora de l'objectiu), aquestes s'han dinterpretar amb precaució, i s'ha de consultar als professionals mèdics per a les decisions clíniques.
  Mitigation Strategy: Supervisió de lús si mai es fa, tot i que està destinat a objectius acadèmics







Per a vaure el model_card millor, dirigir-se al final del notebook. Per a referència, els apartats equivalents a aquesta documentació tenen nom similar o idèntic al notebook.