Informe Pràctica Cirrhosis

Introducció a l'Aprenentatge Automàtic

Jaume Mora i Ladària

Grau en Intel·ligència Artificial
Universitat Politècnica de Catalunya
Barcelona, a desembre de 2023

Índex

Introducció

En aquest document podem trobar la documentació referent a les descripcions i justificacions de la implementació que he realitzat així com dels resultats que he obtingut a la pràctica de Cirrosi realitzada per en Jaume Mora per a l'assignatura d'Introducció a l'Aprenentatge Automàtic.

L'objectiu principal d'aquesta pràctica de laboratori ha estat familiaritzar-se amb els diferents models KNN, DecisionTree i SVM així com el tractament d'un dataset i les seves dades.

Al problema a resoldre se'ns demana que, mitjançant els tres models esmentats anteriorment, fem una predicció de la variable Status que determina si un pacient malalt de cirrosi ha sobreviscut, ha sobreviscut gràcies a un trasplantament de fetge o ha mort.

Cal esmentar que ja que jo m'he canviat de carrera aquest any, venint d'Enginyeria Informàtica no he fet mai encara cap assignatura d'estadística i no ha sigut fàcil per mi tots els apartats d'avaluar dades, tot i així m'ha agradat aprendre'n i crec que he obtingut un molt bon resultat.

Finalment, cal afegir que totes les seccions del notebook es corresponen meticulosament amb les seccions de la documentació.

INFORMACIÓ IMPORTANT SOBRE MODELCARDTOOLKIT:

El notebook consta d'una part d'imports i instal·lacions pel model card així com una secció sencera pel Model Card. L'he entregat executat i per tant s'hauria de veure, de totes maneres és possible que al donar-li a executar al notebook falli la part del ModelCard Toolkit. De ser així, hi ha els passos a seguir al notebook.

1. Anàlisi i preprocessat de dades

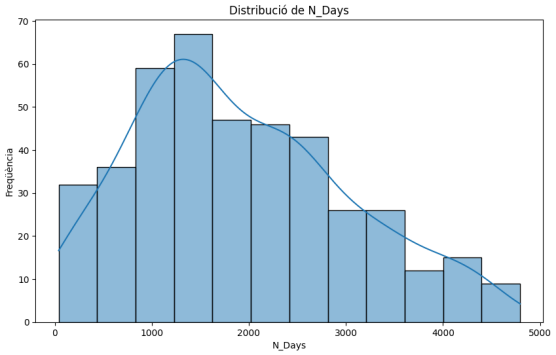
1.1. Anàlisi estadístic de les variables

Fem un anàlisi estadístic de totes les variables. La variable ID no l'he inclòs a l'anàlisi estadístic perquè ja he vist que és simplement un identificador que l'únic que fa és aportar soroll i a més tots els valors d'ID són únics.

Variable N_days:

Anàlisi de la variable N_Days

	count	mean	std	min	25%	50%	75%	max
N_Days	418	1917.78	1104.67	41	1092.75	1730	2613.5	4795



Imatges de l'anàlisi i distribució de la variable

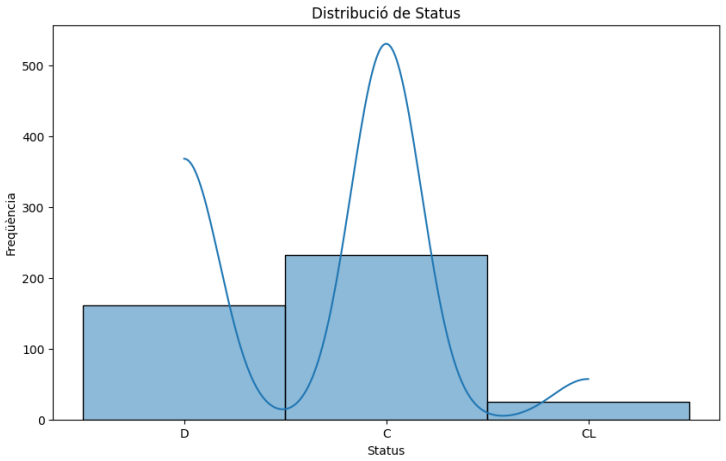
Amb 418 valors, la variable *N_Days* presenta una distribució amb una mitjana de gairebé 1918 dies, però amb una gran desviació estàndard superior a 1100 dies. Això explica que malgrat hi ha una tendència hi ha bastantes dades més individuals

Les dades entre 1000 i 2000 dies indiquen el període típic per a la variable N_days mentre que els valors més extrems podrien representar casos especials o rars. L'histograma, amb la seva forma asimètrica, confirma aquesta interpretació, mostrant que la majoria de dades cauen en aquest rang típic però amb algunes excepcions.

Variable Status:

Anàlisi de la variable Status

	count	unique	top	freq
Status	418	3	C	232



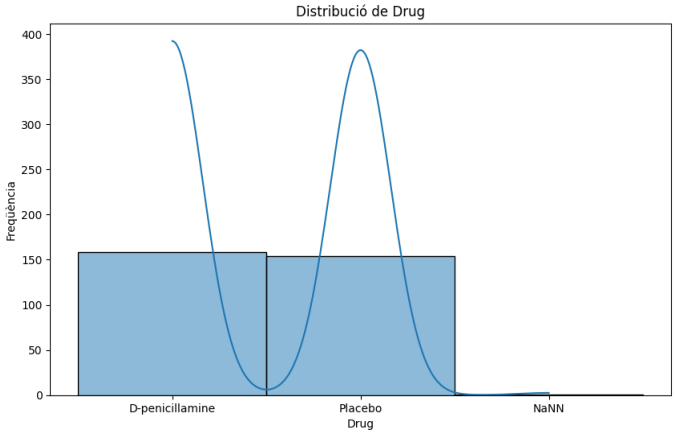
Imatges de l'anàlisi i distribució de la variable

Per a la variable Status veiem clarament com tenim només tres estats possibles: mort, sobreviscut i sobreviscut gràcies a un transplantament de fetge. Aquesta serà la variable a predir però ja veiem que, en aquest dataset, la majoria dels pacients han sobreviscut.

Variable Drug:

Anàlisi de la variable Drug

	count	unique	top	freq
Drug	313	3	D-penicillamine	158



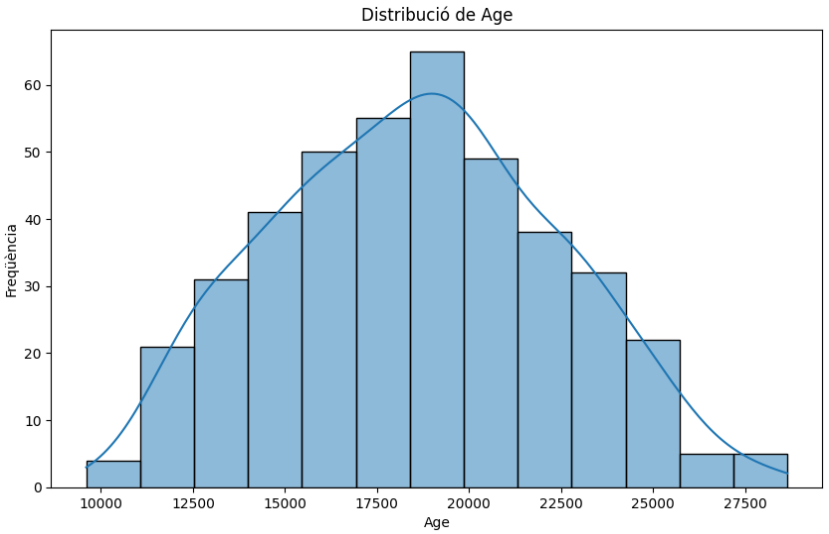
Imatges de l'anàlisi i distribució de la variable

L'anàlisi de la variable "Drug" mostra que totes les dades estan dividides entre dos tipus de Drugs, exceptuant alguns missings. Com ja he dit, la presència dels missings ens suposarà un problema més tard i farà que els haguem d'imputar. Encara ho he d'acabar d'estudiar però tindrem dos opcions, o bé imputar les dades o també plantejar-se esborrar aquesta variable ja que podria ser bastant sorollosa, cosa que estudiaré més endavant.

Variable Age:

Anàlisi de la variable Age

	count	mean	std	min	25%	50%	75%	max
Age	418	18533.4	3815.85	9598	15644.5	18628	21272.5	28650



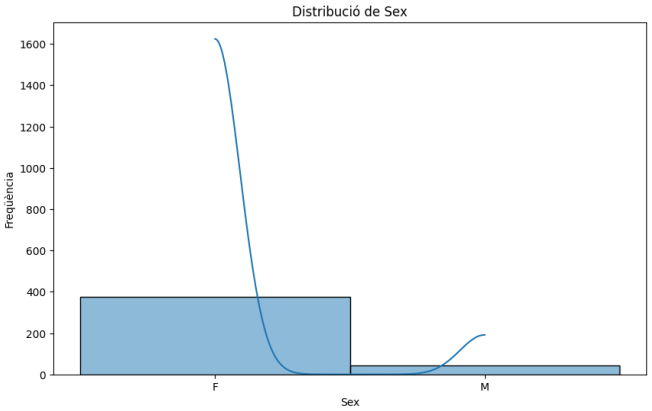
Imatges de l'anàlisi i distribució de la variable

Sabem que l'edat està expressada en dies, per tant la mitjana d'edat que és d'uns 18500 dies equival a aproximadament 50 anys. A més a més, veiem com la desviació estàndard és d'uns 10,5 anys. La distribució de l'edat, com podem veure reflectit en el gràfic, explica que una gran majoria està al voltant dels 50 anys, amb una disminució notable de mostres a mesura que l'edat disminueix i augmenta. Els percentils indiquen que el 25% dels participants tenen menys de 43 anys, el 50% tenen menys de 51 anys, i el 75% tenen menys de 60 anys.

Variable Sex:

Anàlisi de la variable Sex

	count	unique	top	freq
Sex	418	2	F	374



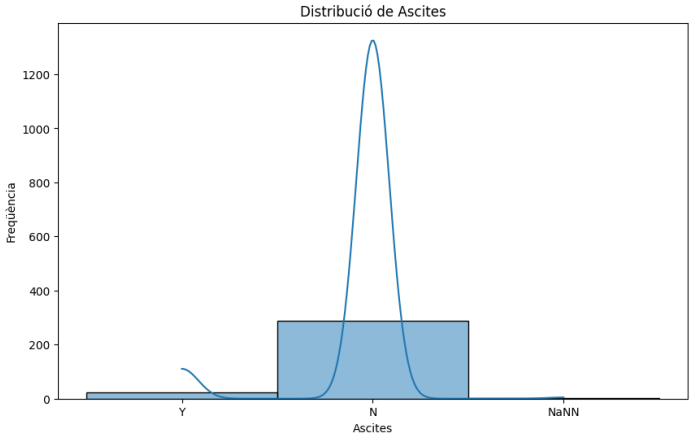
Imatges de l'anàlisi i distribució de la variable

La dada Sex està clarament en desequilibri. Tenim que 374 de 418 dades són F (participants femenines) cosa que ens evidencia que tenim molts pocs M. Això ens podrà comportar alguns problemes importants alhora de predir la variable status més tard. Potser ens ajudarà utilitzar mètodes de ponderació per ajustar per aquest desequilibri.

Variable Ascites:

Anàlisi de la variable Ascites

	count	unique	top	freq
Ascites	313	3	N	288



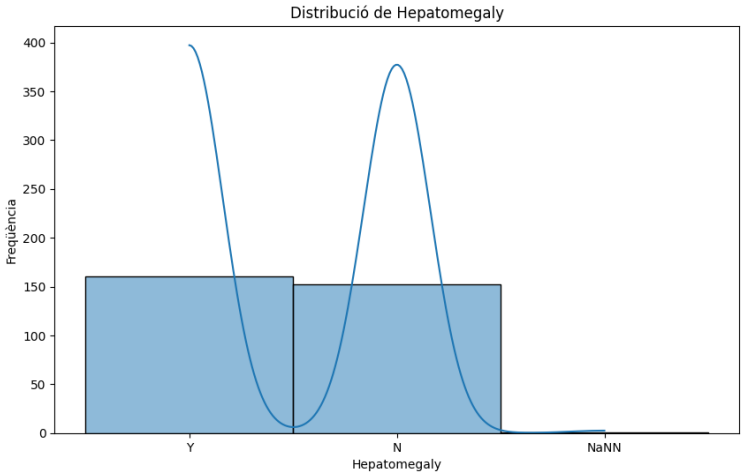
Imatges de l'anàlisi i distribució de la variable

La variable Ascites ens indica la presència de líquid a l'abdomen. Tenim 313 dades reals i, per tant, tota la resta són missings. La categoria més freqüent és "N", que indica que no es té Ascites. El gràfic mostra una distribució molt desequilibrada, amb un nombre significativament més alt de pacients sense ascites. Un cop més, aquest desequilibri podria afectar posteriorment en l'aplicació dels models.

Variable Hepatomegaly:

Anàlisi de la variable Hepatomegaly

	count	unique	top	freq
Hepatomegaly	313	3	Y	160



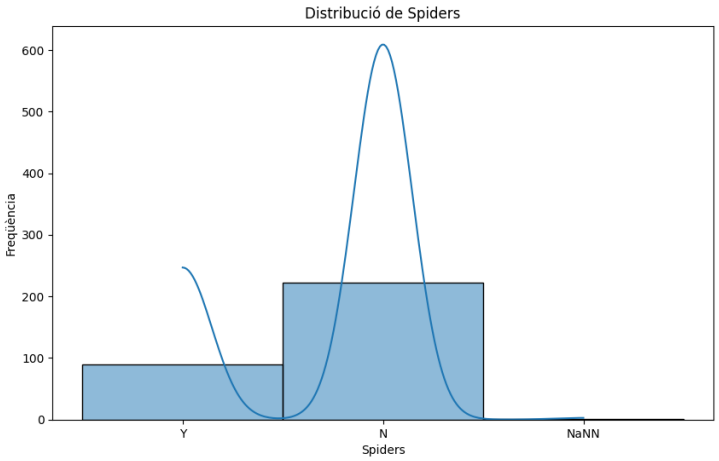
Imatges de l'anàlisi i distribució de la variable

“Hepatomegaly” indica si als pacients els ha augmentat el fetge. Veiem que en aquest cas es mostra un equilibri més proper entre les dues categories en comparació amb altres variables com Ascites. Així i tot, aquesta variable també compta amb missings que haurem d’imputar posteriorment.

Variable Spiders:

Anàlisi de la variable Spiders

	count	unique	top	freq
Spiders	313	3	N	222



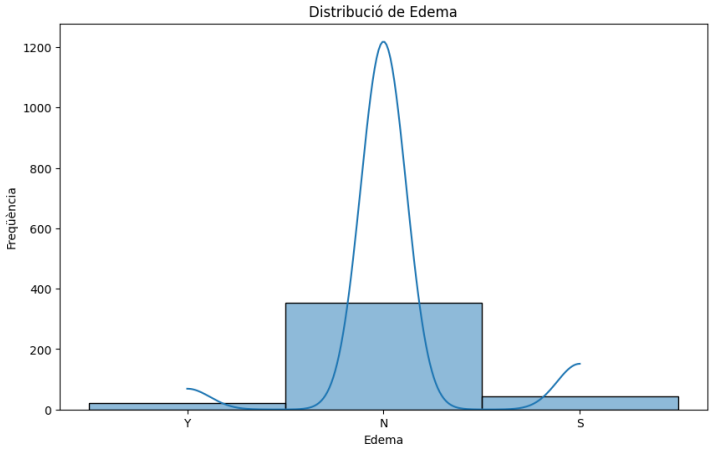
Imatges de l’anàlisi i distribució de la variable

La variable "Spiders" mostra una predominança de la categoria 'N', suggerint que la majoria dels enquestats no tenen Spiders. Veiem que les categories no estan gaire equilibrades ja que ‘N’ duplica a ‘Y’. A més a més, veiem com tenim missings que necessitaran imputació en l'anàlisi posterior.

Variable Edema:

Anàlisi de la variable Edema

	count	unique	top	freq
Edema	418	3	N	354



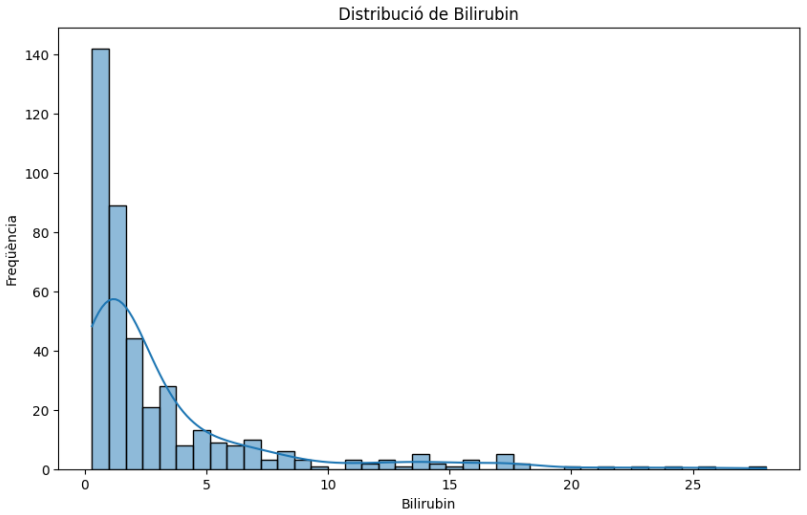
Imatges de l'anàlisi i distribució de la variable

"Edema" reflecteix l'estat de l'edema en els pacients, on 'N' representa absència d'edema sense tractament amb diürètics, 'S' indica edema resolt amb diürètics, i 'Y' mostra edema malgrat la teràpia diürètica. Segons l'anàlisi, la majoria dels pacients no presenta edema. La distribució de la variable "Edema" no mostra missings però, en canvi veiem un desequilibri entre les categories.

Variable Bilirubin:

Anàlisi de la variable Bilirubin

	count	mean	std	min	25%	50%	75%	max
Bilirubin	418	3.22081	4.40751	0.3	0.8	1.4	3.4	28



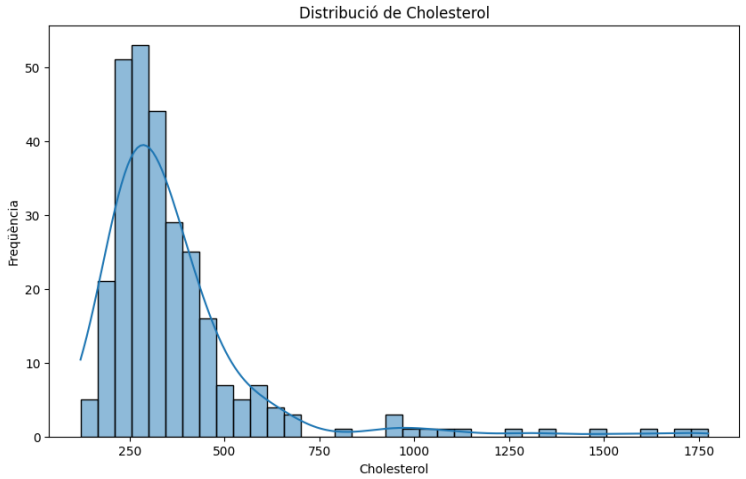
Imatges de l'anàlisi i distribució de la variable

L'anàlisi de la variable "Bilirubin" indica els nivells de bilirubina en els pacients. Els resultats mostren que, en aquesta mostra de 418 pacients (sense missings, per tant), la mitjana de bilirubina és de 3.22, amb una desviació estàndard de 4.41, el que suggereix una variabilitat significativa. Els nivells mínims són de 0.3, i els màxims de 28. Veiem com sembla ser que tenim outliers que haurem de tractar posteriorment.

Variable Cholesterol:

Anàlisi de la variable Cholesterol

	count	mean	std	min	25%	50%	75%	max
Cholesterol	284	369.511	231.945	120	249.5	309.5	400	1775



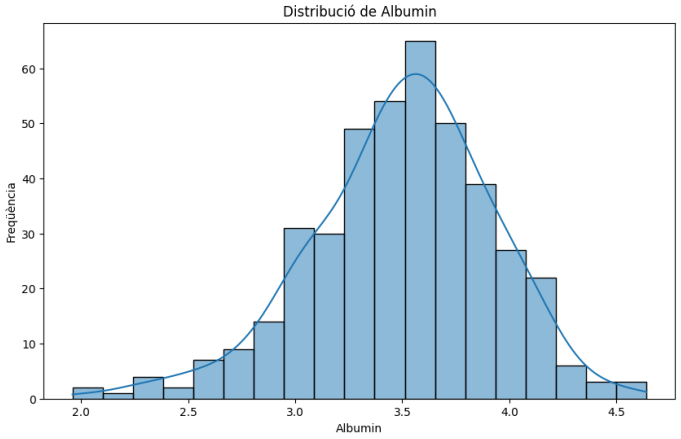
Imatges de l'anàlisi i distribució de la variable

La variable "Cholesterol" quantifica els nivells de colesterol en la sang dels pacients. Amb una mostra de 284 individus i per tant amb bastants missings, la mitjana de colesterol és de 369.5, amb una variació considerable com ho mostra la desviació estàndard de 231.9. Els valors mínims i màxims són de 120 i 1775 demostrant una gran diferència. El gràfic de distribució mostra que la majoria dels nivells de colesterol estan al voltant de la mitjana tot i que sembla que tenim presencia d'outliers, nivells de Cholesterol molt alts que se surten bastant de la tendència. Els percentils mostren que el 25% dels pacients tenen nivells inferiors a 249.5, la mediana està en 309.5, i el 75% tenen nivells inferiors a 400.

Variable Albumin:

Anàlisi de la variable Albumin

	count	mean	std	min	25%	50%	75%	max
Albumin	418	3.49744	0.424972	1.96	3.2425	3.53	3.77	4.64



Imatges de l'anàlisi i distribució de la variable

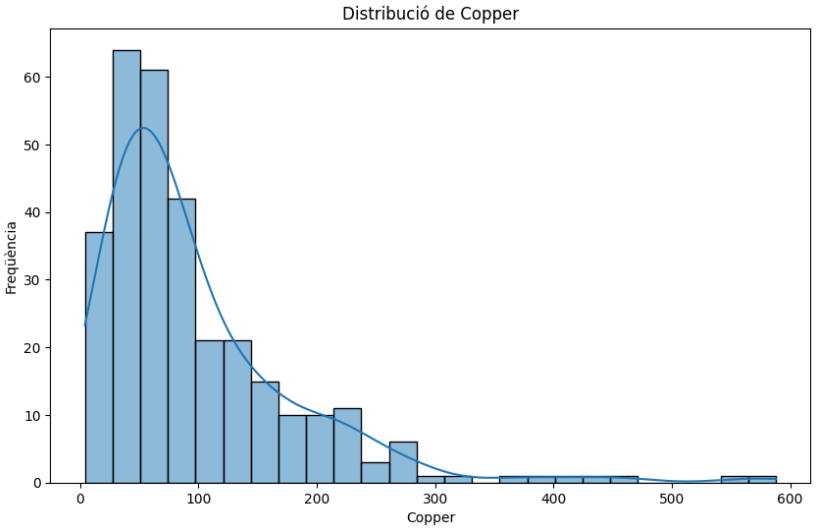
L'Albumin és una variable numèrica amb 418 dades i, per tant, sense missings. La mitjana d'albumina és d'uns 3.50 amb una desviació estàndard de 0.425. El valor mínim registrat és de 1.96, mentre que el valor màxim és de 4.64.

Els percentils mostren una distribució on el 25% dels valors són inferiors a 3.2425, la mediana (el 50% dels valors) és d'aproximadament 3.53, i el 75% dels valors són inferiors a 3.77. Podem concloure, per tant, que aquesta dada no conté outliers.

Variable Copper:

Anàlisi de la variable Copper

	count	mean	std	min	25%	50%	75%	max
Copper	310	97.6484	85.6139	4	41.25	73	123	588



Imatges de l'anàlisi i distribució de la variable

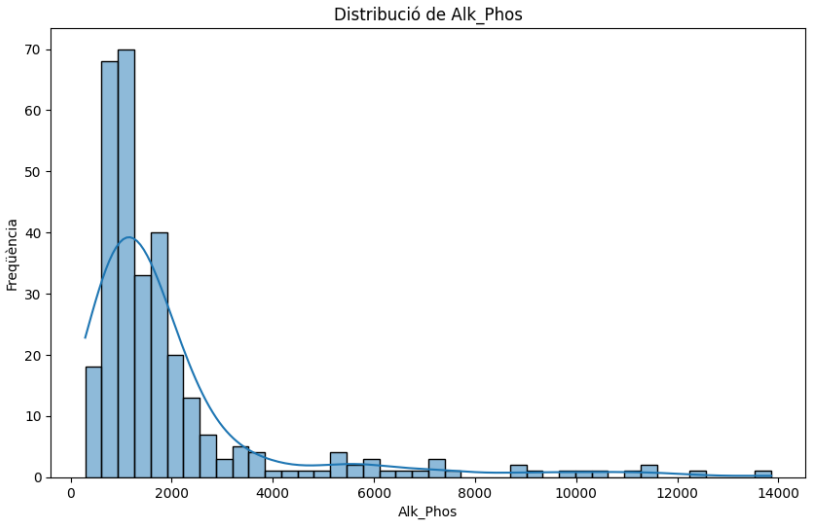
La variable Copper té 310 valors, el que demostra que conté bastants missings. Veiem com la mitjana és d'uns 97.6484 micrograms per decilitre i que té una gran desviació estàndard de 85.6139. El valor mínim és de 4 i el màxim és de 588. És per això i també veient la distribució que podem afirmar la presència d'outliers en aquesta variable Copper.

Els percentils mostren que el 25% dels valors són inferiors a 41.25, la mediana és de 73, i el 75% són inferiors a 123.

Variable Alk_Phos:

Anàlisi de la variable Alk_Phos

	count	mean	std	min	25%	50%	75%	max
Alk_Phos	312	1982.66	2140.39	289	871.5	1259	1980	13862.4



Imatges de l'anàlisi i distribució de la variable

La variable Alk Phos té 312 valors per tant sabem que conté un centenar de missings. La seva mitjana és d'uns 1983. La desviació estàndard és bastant alta, d'uns 2140.39. El valor mínim mesurat és de 289, mentre que el màxim arriba a 13862. Això i la distribució del gràfic ens demostren perfectament que tenim bastants outliers d'aquesta variable.

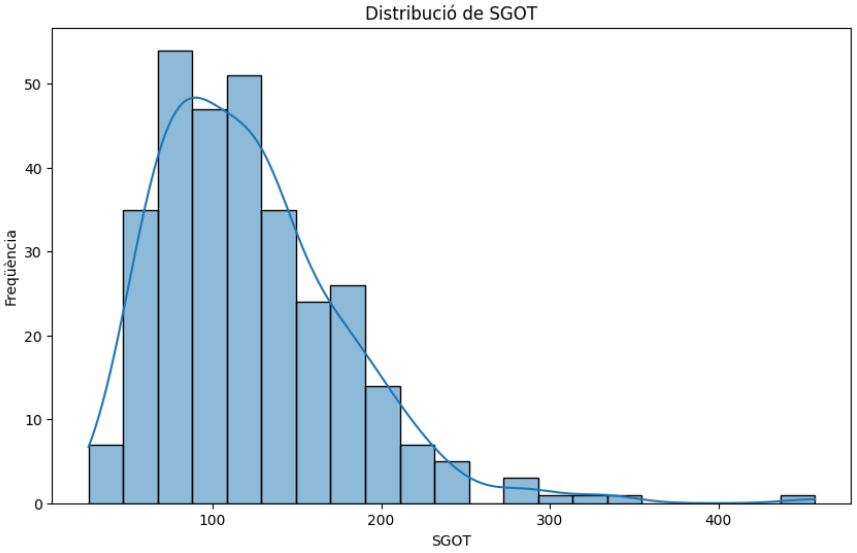
Els percentils indiquen que el 25% dels valors estan per sota de 871.5, la mediana (el 50% dels valors) és de 1259, i el 75% són per sota de 1980.

La gran majoria de les dades dades de la gràfica estan concentrades en valors més baixos, tot i que tenim uns outliers amb valors excepcionalment alts.

Variable SGOT:

Anàlisi de la variable SGOT

	count	mean	std	min	25%	50%	75%	max
SGOT	312	122.556	56.6995	26.35	80.6	114.7	151.9	457.25



Imatges de l'anàlisi i distribució de la variable

La variable SGOT consta de 312 dades i, per tant, un altre centenar de missings. La mitjana de SGOT és de 122.556 amb una desviació estàndard, prou alta, de 56.7. El valor mínim és de 26.35 mentre que el màxim és de 457.25.

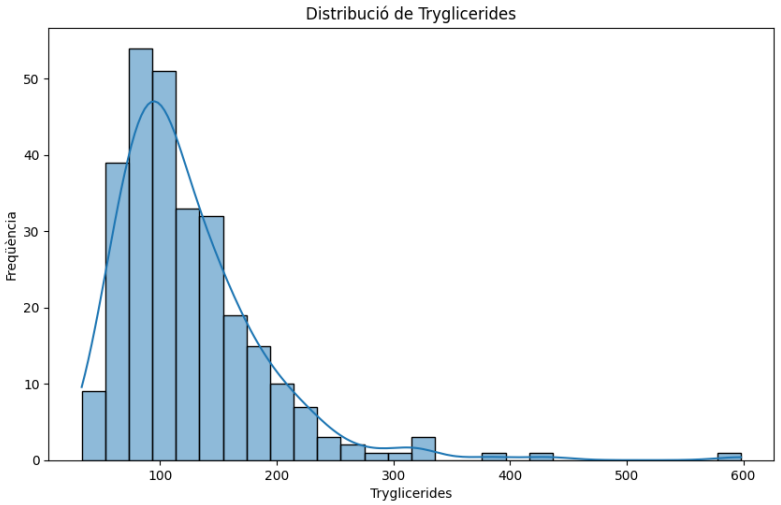
Els percentils reflecteixen que el 25% dels valors estan per sota de 80.6, la mediana és de 114.7, i el 75% estan per sota de 151.9.

Amb tot això i amb la distribució al gràfic podem afirmar que comptem amb la presència d'outliers.

Variable Tryglicerides:

Anàlisi de la variable Tryglicerides

	count	mean	std	min	25%	50%	75%	max
Tryglicerides	282	124.702	65.1486	33	84.25	108	151	598



Imatges de l'anàlisi i distribució de la variable

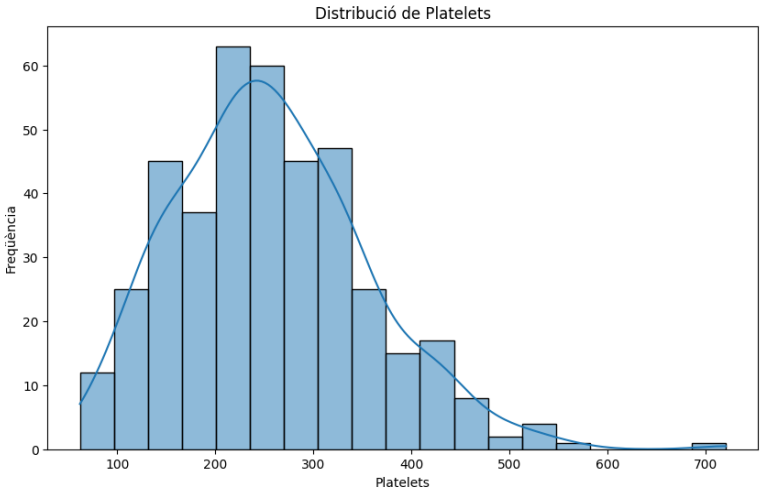
La variable SGOT compta amb 312 dades, la qual cosa indica la presència d'un altre centenar de missings. La mitjana de SGOT és de 122.556, amb una desviació estàndard alta de 56.7, el que ens mostra una variabilitat considerable en les dades. El valor mínim registrat és de 26.35, en contrast amb el valor màxim que ascendeix fins a 457.25. Amb aquests valors i fixant-nos en la distribució del gràfic podem afirmar que tenim outliers.

En aquest cas, els percentils ens diuen que el 25% dels valors són inferiors a 80.6, la mediana és de 114.7 i el 75% dels valors són menors de 151.9.

Variable Platelets:

Anàlisi de la variable Platelets

	count	mean	std	min	25%	50%	75%	max
Platelets	407	257.025	98.3256	62	188.5	251	318	721



Imatges de l'anàlisi i distribució de la variable

La variable Platelets presenta un total de 407 dades, per tant, tot i que conté missings en té molt pocs. Veiem com la mitjana és de 257.025, amb una desviació estàndard de 98.3256. El valor mínim és notablement baix (62) comparat amb el valor màxim (721).

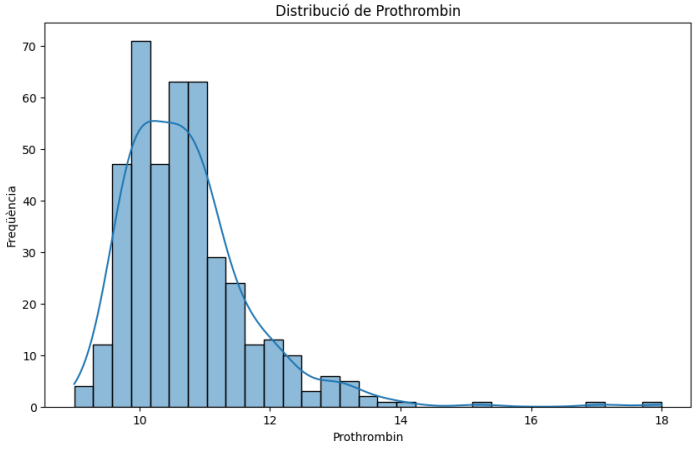
Analitzant els percentils, podem veure que el 25% dels valors són inferiors a 188.5, la mediana és de 251 i, finalment, el 75% dels valors són inferiors a 31.

Amb la distribució mostrada al gràfic es nota una certa asimetria a la dreta. Això corrobora la presència d'uns pocs outliers.

Variable Prothrombin:

Anàlisi de la variable Prothrombin

	count	mean	std	min	25%	50%	75%	max
Prothrombin	416	10.7317	1.022	9	10	10.6	11.1	18



Imatges de l'anàlisi i distribució de la variable

L'anàlisi de la variable Prothrombin mostra que tenim 416 dades per tant, en aquest cas no tenim outliers. La mitjana dels nivells de Prothrombin és de 10.7317, amb una desviació estàndard prou baixa de 1.022, indicant que les dades estan bastant agrupades al voltant de la mitjana.

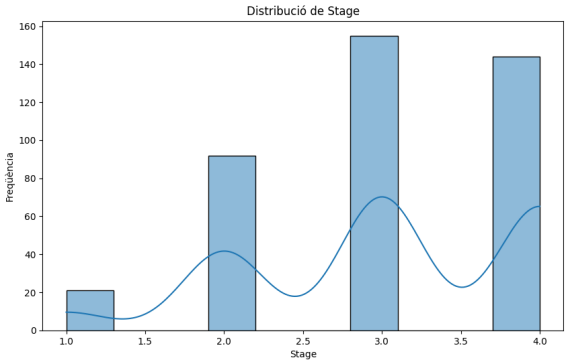
El valor mínim registrat per Prothrombin és de 9 mentre que el màxim és, el doble. Aquesta diferència i la distribució del gràfic en el qual veiem algunes dades perdudes a la dreta ens confirmen la presència d'outliers.

Els percentils mostren que el 25% de les dades són inferiors a 10, la mediana és molt propera a la mitjana amb un valor de 10.6, i el 75% de les dades són inferiors a 11.1.

Variable Stage:

Anàlisi de la variable Stage

	count	mean	std	min	25%	50%	75%	max
Stage	412	3.02427	0.882042	1	2	3	4	4



Imatges de l'anàlisi i distribució de la variable

La variable Stage, que és categòrica, defineix en quin estat està el pacient. Veiem que tenim una gran majoria de pacients en estat 3 i 4. Els altres, tot i que són bastant menys estan distribuïts en els estats 1 i 2. Tenim 412 dades per tant, veiem que tenim molts pocs missings.

1.2. Estudi de balanceig de classes

Aquest punt en realitat l'he estudiat a l'apartat anterior. Vist que sí que tenim un desequilibri més que notable en algunes de les variables, he optat per intentar aplicar un SMOTE. Un SMOTE és un mètode d'oversampling que soluciona el desequilibri de classes generant mostres sintètiques per a la classe minoritària.

De totess maneres, no acabo d'entendre perquè l'SMOTE fa que em surtin resultats pitjors als que tenia sense aplicar-lo.

És a dir, tot i estudiar la possibilitat d'aplicar un SMOTE, he considerat que no era una bona opció, ja que semblava empitjorar el rendiment dels models. Potser altres factors específics del conjunt de dades o del problema podrien estar influint en aquest resultat contradictori, però tot i estudiar-ho no he pogut treure'n l'entrellat.

1.3. Missings: Identificació i gestió

Veiem a les següent taules com, efectivament, tenim molts missings. És per això que ens cal fer un bon estudi dels missings i eliminar-los o imputar-los si cal.

Nom de la variable	Drug	Ascites	Hepatomegaly	Spiders	Cholesterol	Copper
Nombre de missings	86	86	86	86	103	88

Nom de la variable	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
Nombre de missings	86	86	105	8	1	5

Recordem que el tractament dels missings ha estat posterior a la partició de train i test.

Per a la imputació de les dades categòriques (que tot i tenir-les de forma numèrica segueixen sent categòriques) he optat per utilitzar un SimpleImputer de SKLearn utilitzant l'estratègia most_frequent. Això fa que, per a les variables categòriques a imputar, se substitueixen els missings per la moda. Havia pensat diferents opcions per a les variables categòriques i, realment, ha sigut difícil trobar la millor manera d'imputar-les ja que tot i tenir-les de forma numèrica no podiem deixar d'entendre que eren catagòriques. És a dir que si els valors eren 0 i 1, els valors imputats havien de ser 0 o 1, no podien ser 0.6, per exemple. És per això que he arribat a la conclusió que la millor manera era fer un imputador amb la moda.

Per a les variables numèriques he utilitzat un KNNImputer. Per a cada missing, aquest imputador troba els K veïns més propers (on K, en aquest cas la millor opció és que sigui 5), i substitueix aquest missing per la mitjana dels veïns.

Tant per al SimpleImputer per a les categòriques com per al KNNImputer de les numèriques he fet fit amb el train i he fet transform al train i al test.

Un cop aplicats els imputadors he comprovat quants missings tenia a les dades. Com que el resultat ha estat 0 significa que he imputat correctament els missings i que, per tant, ara ja tinc unes dades netes de missings.

1.4. Outliers: Identificació i gestió

Com hem vist també abans quan fèiem l'anàlisi de les variables tenim bastantes variables amb outliers. De fet per algunes d'elles els outliers representen gairebé un 25% dels valors.

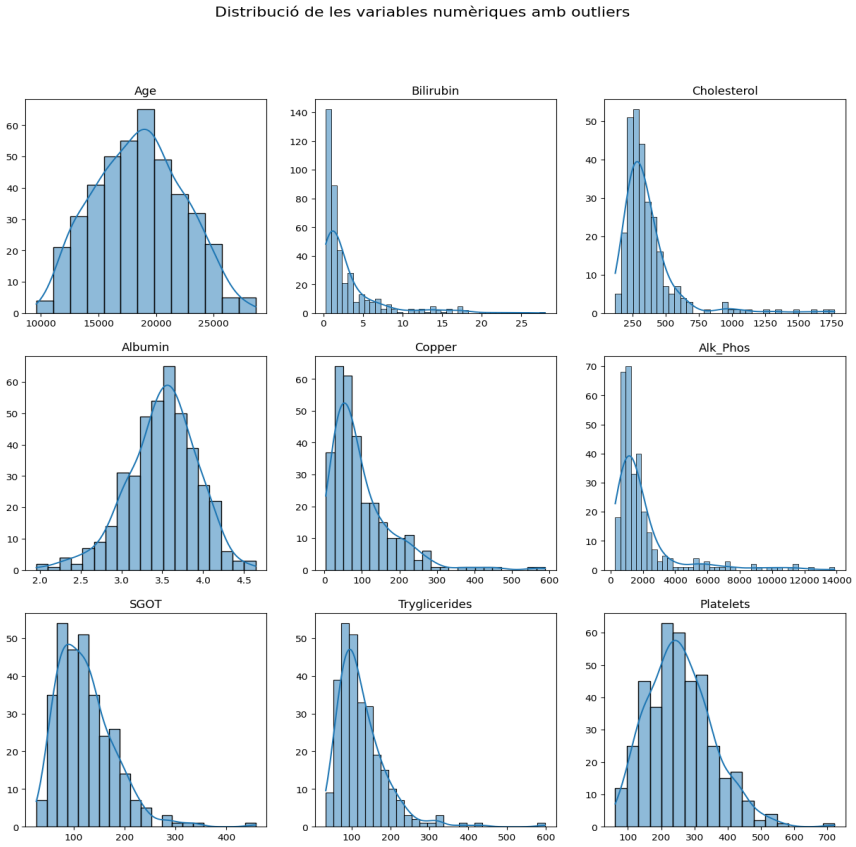
Com que veia que els outliers em corrompien molt el rendiment dels models primer he optat per imputar-los.

No penso que sigui una bona opció eliminar-los, ja que tenim un dataset molt petit, amb poc més de 400 dades i eliminant els outliers ens queda un dataset amb menys de 300 dades. Per això, he acabat optant per passar els outliers a missings per poder-los imputar posteriorment.

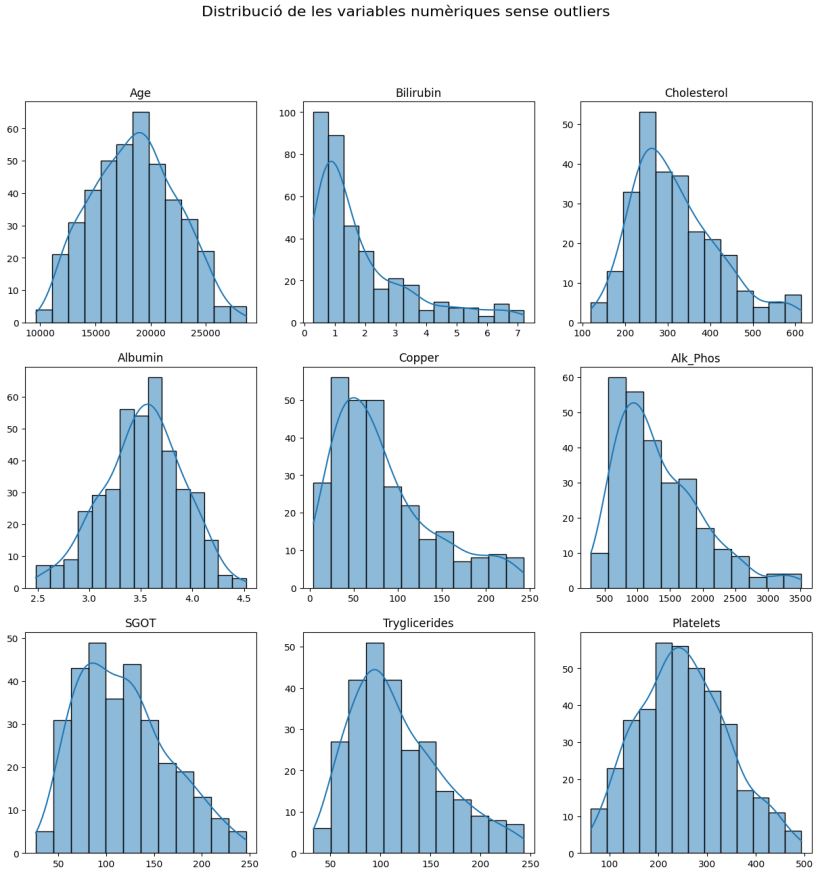
Recordo que els outliers els he tractat abans de fer la partició de train i test.

El tractament dels outliers l'he fet posant un límit inferior i superior a cada una de les dades. Aquest límit l'he calculat utilitzant l'IQR (interquartile range). Recordem que el que fa és restar el tercer quartil amb el primer i multiplicar el resultat per 1.5.

Llavors assumim que tots els valors que se sobresurten d'aquests paràmetres són outliers i per tant els convertim a missings.



Imatge de les variables numèriques amb outliers.



Imatge de les variables numèriques sense outliers.

1.5. Recodificació de variables

He recodificat totes les variables categòriques passant-les a numèriques a mà. Per a aquelles amb dos valors he assignat 0 i 1 i a totes les altres he assignat un número entre 0 i 1 proporcionalment amb el nombre de valors diferents que tenia. Només per a la variable Status, que és la variable a predir li he assignat els números 0, 1 i 2 perquè sinó més tard em donava problemes al passar aquesta variable a integer (es convertia tot a 0 o 1).

1.6. Particionat del dataset

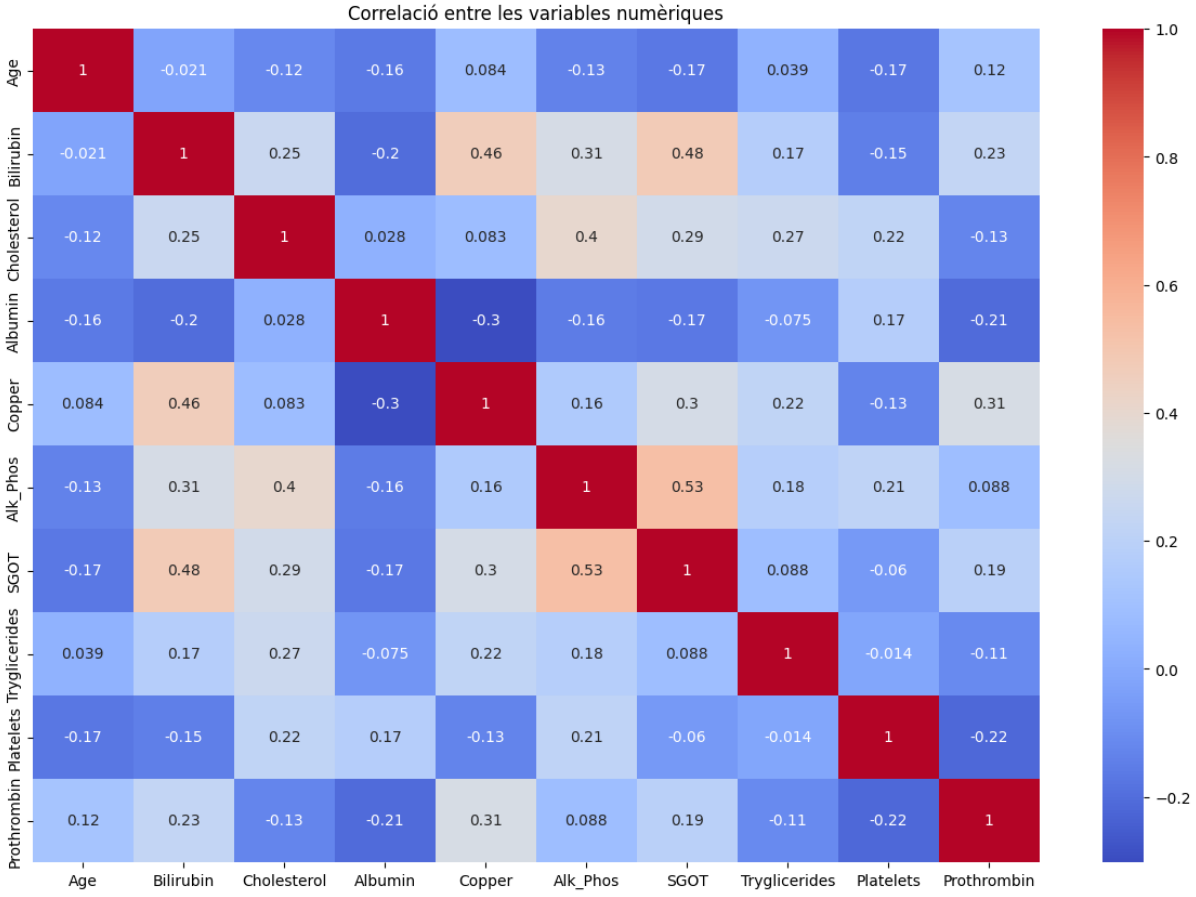
Per fer el particionat del dataset he observat que la millor distribució era dedicar un 70% de les dades per al train i el 30% restant per al test. No he considerat necessari utilitzar partició de validation perquè he utilitzat després Grid Search i Cross Validation.

2. Preparació de variables

2.1. Normalització de variables

Abans d'imputar els missings he volgut escalar totes les dades entre 0 i 1. Per a les numèriques he vist que la millor opció era utilitzar un MinMaxScaler per escalar-les totes de 0 a 1. Per a les categòriques he convertit jo, manualment, totes les dades a numèriques. És a dir, per exemple a les variables amb tres categories he utilitzat 0 per a la categoria A, 0.5 per a la categoria B i 1 per a la categoria C. Tot això m'ha ajudat molt per donar-li el mateix pes a totes les dades i per tenir-les escalades durant tota la pràctica. Tenia el problema, sinó, que utilitzant depèn de quins models a algunes variables se li donaven més pes que a d'altres.

2.2. Anàlisi de correlacions entre variables numèriques



La matriu de correlació mostra la relació de variables numèriques amb valors des de -1 a 1. Quan el valor és 1 significa una correlació positiva perfecta (quan augmenta un augmenta l'altre), quan el valor és 0 indica que no hi ha correlació i quan el valor és -1 indica una correlació negativa perfecta (quan una variable augmenta l'altra disminueix).

Per tant, les variables amb una correlació més notable són:

- **Bilirubin i SGOT:** Amb una correlació positiva del 0.48. Segons sembla, quan la bilirrubina augmenta també ho fa l'SGOT.
- **Albumin i Copper:** Amb una correlació negativa del -0.3 explica que quan una de les dues variables augmenta l'altra tendeix a disminuir. Veiem com en aquest cas la correlació no és que sigui molt forta.
- **Alk Phos i SGOT:** Existeix una correlació positiva prou alta (0.53) entre relativament forta Alk Phos i SGO. El que suggereix que ambdues variables augmenten juntes.

- **Platelets i Prothrombin:** Es mostra una correlació negativa de -0.22. Tot i no ser molt alta la correlació sembla que, en alguns casos, quan disminueix una augmenta l'altra..
- **Copper i Bilirubin:** Amb una correlació prou alta de 0.46, aquesta relació positiva suggereix que quan augmenta Copper tendeix a augmentar Bilirubin.

De totes maneres, com que les correlacions no són gaire altes, en aquest pas **he decidit no eliminar cap variable.**

2.3. Anàlisi de variables categòriques i variable objectiu

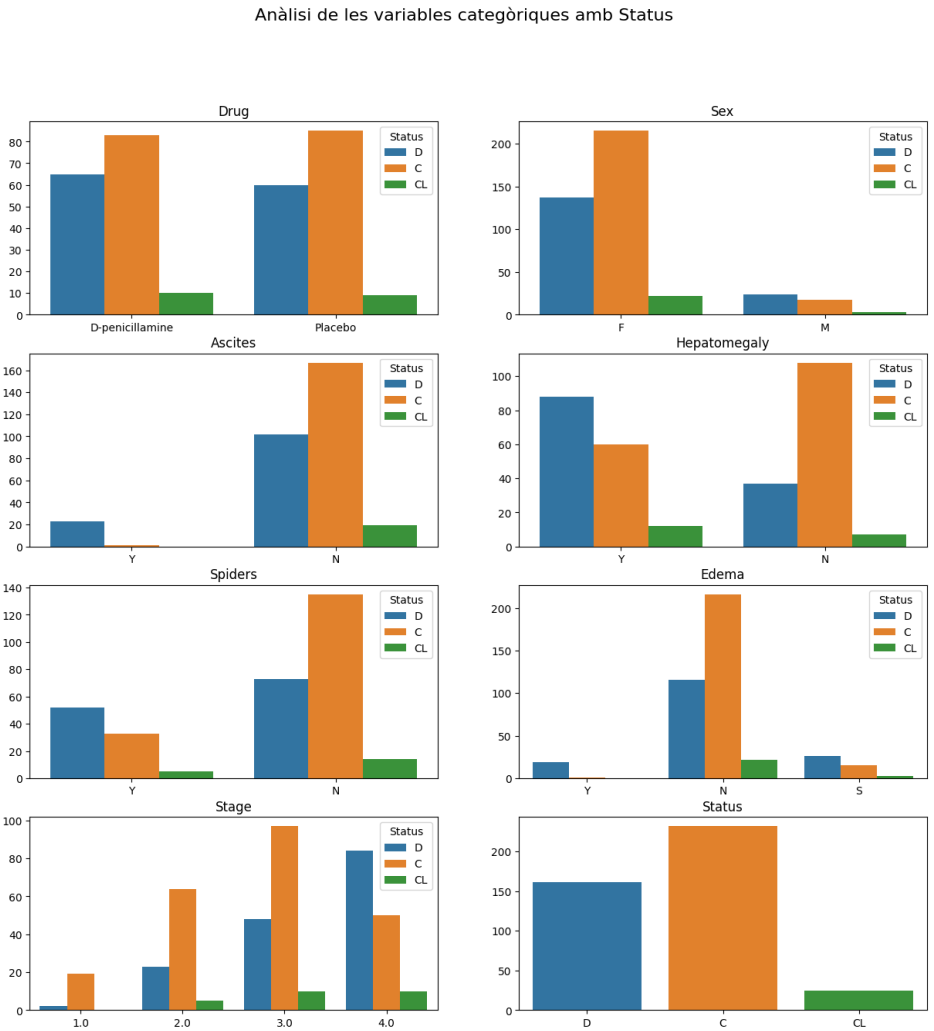


Figura de l'anàlisi bivariada amb les variables categòriques i la variable objectiu.

Drug: Per a la variable Drug podem observar com el fet d'utilitzar una o altra fa que pràcticament no canviï gens el resultat, sent la "Sobreviure" el resultat més freqüent.

Sex: Aquí notem una diferència, a part de veure que tenim més dones. Mentre les dones acostumen a sobreviure, els homes tenen la mort com a classe majoritària.

Ascites: Veiem com la presència d'ascites assegura gairebé la mort. Gairebé tots els pacients amb Ascites acaben morint. En canvi, els que no en tenen solen sobreviure.

Hepatomegaly: Com en el cas d'ascites però amb una diferència menor, la presència d'Hepatomegaly està associada amb una major nombre de morts. Els pacients sense Hepatomegaly sobreviuen normalment.
Spiders: La presència de "Spiders" segueix la mateixa tendència que les variables "Ascites" i "Hepatomegaly". Veiem com proporcionalment tenim molts més morts en els pacients amb Spiders que no pas sense.

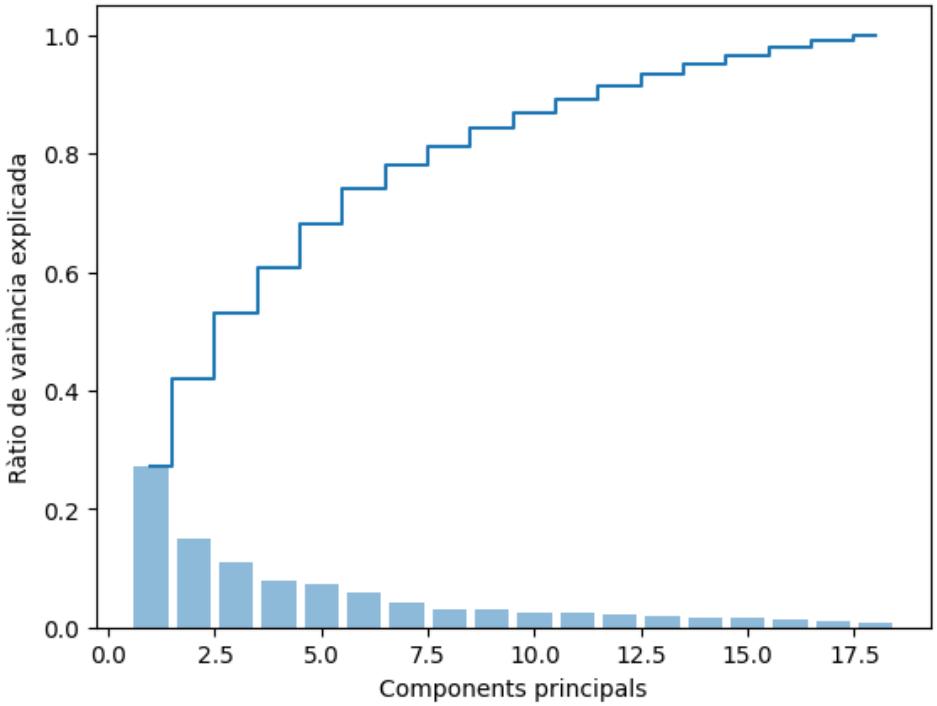
Edema: Els pacients sense edema (N) tendeixen a sobreviure sense necessitat de trasplantament. En canvi, aquells amb edema tenen gairebé la mort assegurada. Veiem com la majoria, però no tots els que tenen edema sense diuretics (S) moren.

Stage: Veiem que, en proporció, a mesura que l'etapa de la malaltia avança la mortalitat augmenta.

2.4. Eliminació de variables redundants o sorolloses

A part d'haver eliminat la variable 'ID' des del principi perquè ja he vist que no aportava cap mena d'informació, ara me n'he adonat de que també val la pena eliminar la variable **'Drug'** perquè no ens aporta res. Si ens fixem en l'estudi de l'apartat 2.3, tant si es fa servir D-penicillamine com si es fa servir Placebo els resultats a Status són els mateixos.

2.5. Estudi de dimensionalitat amb PCA



Gràfic amb l'estudi de la dimensionalitat amb PCA.

Sí, sembla necessari reduir variables basant-nos en l'anàlisi de PCA de la gràfica.

Si ens hi fixem, els primers components tenen molta variància explicada per a cada component principal, cosa que indica que aporten una quantitat significativa d'informació sobre les dades originals. De totes maneres, mentre anem afegint components principals la variància explicada disminueix, és a dir aquests components aporten menys informació.

Per saber amb quants components quedar-nos podem utilitzar la regla del colze. Consisteix a posar un nombre límit de components a partir del qual la variància explicada gairebé no varia. Se li diu colze per la similitud del punt de la corba al punt on hi ha el colze al braç.

Per tant, amb la reducció de variables trobarem un nombre òptim de components principals que ens simplificarà el model sense perdre atributs significatius. Farà al model més eficient i menys vulnerable davant problemes com l'overfitting.

Veig que el punt de colze podria estar sobre el 5, per tant, utilitzant PCA reduiré les dades a 5 components principals.

3. Definició de models

La part de documentació dels models l'he fer de la següent manera. Per a la definició de mètriques i hiperparàmetres he fet un sol apartat i llavors he fet un apartat per a cada model.

3.1. Definició de mètriques i hiperparàmetres

Per a la definició de mètriques dels tres models he utilitzat GridSearchCV de la llibreria sklearn per a trobar les millors mètriques per a cada model. Tot i que fa que trigui una mica més, he volgut fer 100 particions per a provar el GridSearchCV, perquè he vist que en tots els casos era el que em donava millors paràmetres que reflectien després a un millor model.

Aquí tenim les taules amb els valors provats amb el GridSearchCV i els valors triats:

MODELS	VALORS PROVATS	VALORS TRIATS	TEMPS EXECUCIÓ
KNN	{'n_neighbors': [3, 5, 7, 9, 11, 13, 15, 17, 19, 21], 'weights':	Els millors paràmetres per a un KNN són: {'metric': 'manhattan',	6.5s

	['uniform', 'distance'], 'metric': ['euclidean', 'manhattan', 'minkowski']}	'n_neighbors': 19, 'weights': 'distance'} *	
DecisionTree	{'criterion': ['gini', 'entropy'], 'max_depth': [None, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'min_samples_split': [2, 5, 10, 15, 20], 'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]}	Els millors paràmetres per a un DecisionTree són: {'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 8, 'min_samples_split': 2}	2m 11.6s
SVM	{'C': [0.1, 1, 10, 100, 1000], 'kernel': ['linear', 'poly', 'rbf', 'sigmoid'], 'gamma': ['scale', 'auto']}	Els millors paràmetres per a un SVM són: {'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}	25.9s

*** IMPORTANT:** Ja sé que amb una K de 19, es classificarà malament la classe 'CL'. Tot i així, és la millor manera per treure el millor resultat a l'hora de predir, ja que les dades de la classe 'CL' són minoritàries.

KNN:

Veiem com per al KNN les millors combinacions de mètriques són les següents:

N Neighbors	Weights	Metric	Mean Test Score
19	distance	manhattan	0.736667
21	distance	manhattan	0.731667
15	uniform	euclidean	0.726667
15	uniform	minkowski	0.726667
15	distance	manhattan	0.726667

Millors combinacions de mètriques per al KNN.

DecisionTree:

Per al DecisionTree les millors combinacions són aquestes:

Criterion	Max Depth	Min Samples Split	Min Samples Leaf	Mean Test Score
gini	4	2	8	0.775
gini	4	5	8	0.775
gini	4	10	8	0.775
gini	4	15	8	0.775
gini	4	20	8	0.775

Millors combinacions de mètriques per al DecisionTree.

SVM:

Finalment, per al SVM tenim aquests resultats:

C	Kernel	Gamma	Mean Test Score
10	rbf	scale	0.745000
1	rbf	auto	0.745000
10	rbf	auto	0.743333
1	sigmoid	auto	0.741667
10	poly	scale	0.738333

Millors combinacions de mètriques per al SVM.

3.2. Entrenament dels models

Per a l'entrenament he agafat la millor combinació de mètriques per a cada model, que he descobert fent GridSearchCV a l'apartat anterior i he fet fit amb les dades de train que en cada cas suposen un 70% de les dades, com així ho he definit a l'hora de fer la partició.

Després, he fet predict del test i he acabat comparant cada predict amb el seu test corresponent.

3.3. Anàlisi de resultats i iteració

Veiem com per a tots els models tenim uns resultats bastant similars. En quant a la iteració, he fet un CrossValidation amb 100 particions. De totes maneres, per a cadascun dels models, s'han ajustat els hiperparàmetres de manera iterativa utilitzant el Cross Validation. Això ha permès no només ajustar els models de manera més fina, sinó també comprendre millor el seu comportament en diferents condicions i comparar la diferència entre els resultats amb les diferents mètriques.

Els resultats amb els millors hiperparàmetres en cada cas són els següents:

Model	Accuracy	Precision	Recall	F1 Score
KNN	0.753968	0.733552	0.753968	0.729078

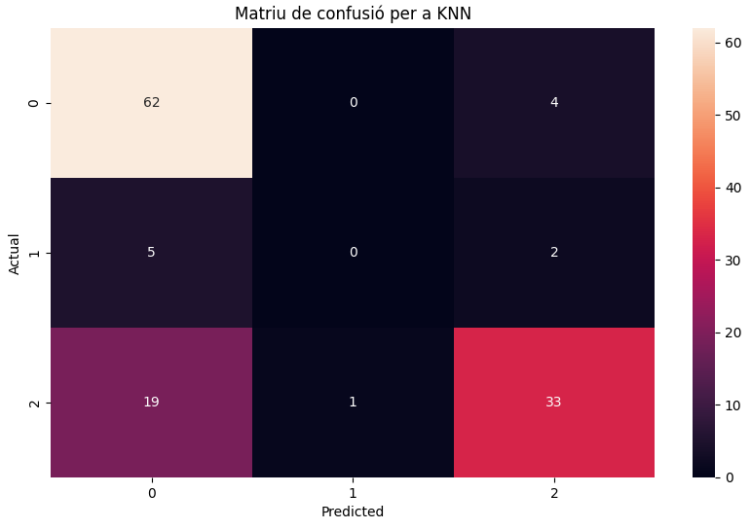
Model	Accuracy	Precision	Recall	F1 Score
DecisionTree	0.730159	0.751091	0.730159	0.705431

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.738095	0.763863	0.738095	0.711612

Resultats dels tres models.

RECORDEM QUE PER A TOTES LES MATRIUS DE CONFUSIÓ: CLASSE 0 ÉS C, CLASSE 1 ÉS CL I CLASSE 2 ÉS D

Matriu de confusió del KNN:



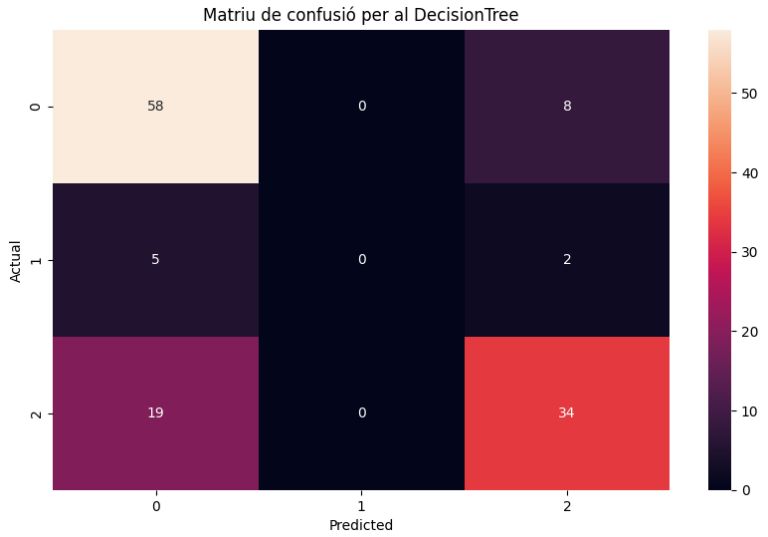
Matriu de confusió del KNN.

La matriu de confusió per al model KNN mostra una distribució desigual en la precisió de la classificació entre les classes. El nombre més gran de prediccions correctes es troba a la classe 0 amb 62 observacions correctament classificades.

De totes maneres, aquesta eficàcia no es reflecteix en les altres classes, hi ha una gran quantitat d'observacions de la classe 2 que han estat classificades incorrectament com si fossin de la classe 0, específicament 19 casos. A més a més, per a la classe 1 només s'ha predit una dada i s'ha fet de forma incorrecta.

Finalment, la predominància de classificacions correctes per a la classe 2 amb 33 encerts indica una capacitat moderada del model per distingir aquesta classe, perquè veiem que també, molts d'ells s'han predit com a classe 0.

Matriu de confusió del DecisionTree:



Matriu de confusió del DecisionTree.

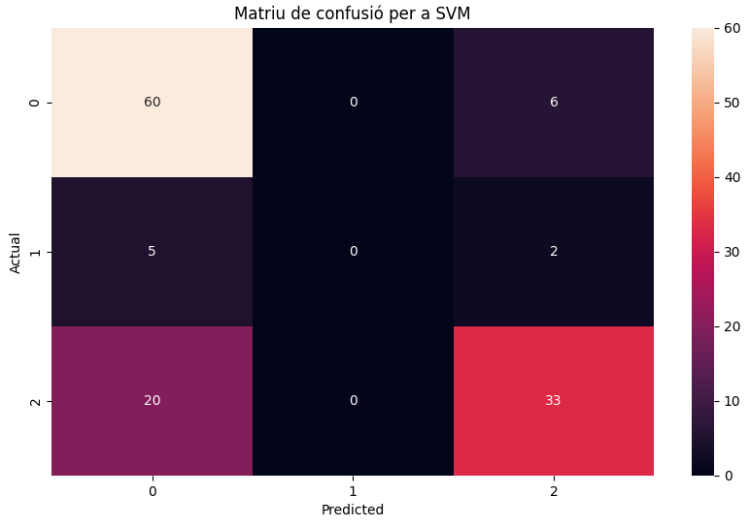
En aquest cas, veiem que el model té un rendiment lleugerament pitjor en comparació amb el KNN.

La classe 0 continua sent la millor predita amb 58 correctes. No obstant això, a diferència de la matriu del KNN, hi ha 8 instàncies que pertanyen realment a la classe 0 però que han estat predites com a classe 2.

Pel que fa a la classe 1, cap instància ha estat predita com a classe 1. Això és degut segurament a una insuficiència de dades de la classe 1.

Per a la classe 2, trobem que hi ha 34 instàncies correctament predites, però també observem que 19 instàncies han estat classificades com a 0.

Matriu de confusió del SVM:



Matriu de confusió del SVM.

El model SVM també ha aconseguit una alta taxa d'encerts en la classe 0. Pel que fa a la classe 1, un cop més veiem que el model no ha predit cap de les seves instàncies ni correcta ni incorrectament.

En la classe 2, el model ha classificat correctament 33 instàncies, però alhora, 20 instàncies de la classe 2 han estat mal identificades com a classe 0. Mantenit-se també en la línia dels altres models.

Conclusió:

Dels tres models, el KNN ha demostrat tenir els millors resultats en aquesta anàlisi particular, encara que no per un marge molt ampli. És important destacar que l'ajust dels hiperparàmetres a través de la validació creuada ha permès millorar la performance dels models i entendre el seu comportament en diferents condicions.

4. Selecció de model

4.1. Descripció del model triat

He triat el model KNN perquè és el que millors resultats em dóna en quant a accuracy, precision i F1 score. Veiem als resultats de l'apartat anterior com els tres models m'han donat uns resultats bastant similars, de totes maneres, per una lleugera diferència el millor ha estat el KNN.

El KNN (K-Nearest Neighbors) és un algoritme que serveix per classificar les dades basant-se en els K elements més propers. És a dir, en aquest cas, com que K és 19,

agafa les 19 dades més properes i fa una votació majoritària. Per a les variables categòriques agafa la classe majoritària en aquestes 19. Per a les numèriques, calcula una mitjana ponderada d'aquestes 19 dades.
A diferència d'altres mètodes, el KNN treballa sobre les dades durant la fase de classificació, cosa que ens porta certs avantatges.

El KNN no és paramètric, per tant no fa suposicions sobre la forma de la distribució de les dades. A més a més, és un algoritme senzill d'entendre i implementar. Amb la tria d'uns bons hiperparàmetres i unes bones dades dóna molts bons resultats.

Totes aquestes característiques a més del bon rendiment del meu KNN han fet que m'hagi decantat per el KNN a l'hora de triar model.

4.2. Anàlisi de les limitacions i capacitats del model

Les limitacions del KNN són diverses i, per això, hem hagut de preparar molt bé les dades abans d'entrenar el model. Algunes de les limitacions més notables són:

- Sensibilitat a la mida del dataset i a la K: el KNN és un mètode altament ineficient per als datasets molt grans o per una gran mida de K, ja que el que fa KNN és calcular totes les distàncies.

- Sensibilitat a les variables irrellevants: KNN dóna el mateix pes a totes les variables, per tant és igual com de rellevants siguin.

- Necessitat d'una normalització i/o estandardització de les dades: el KNN necessita que totes les seves dades estiguin normalitzades per a poder aplicar la mesura de distància.

- Sensible als outliers: els outliers poden portar a prediccions incorrectes.

4.3. Resultats en partició de test amb comparació amb train

	Accuracy	Precisió	Recall	F1 score
Train	1.000000	1.000000	1.000000	1.000000
Test	0.753968	0.733552	0.753968	0.729078

Comparació de resultats per al KNN.

En aquest apartat entenc que he de comentar els resultats del test perquè els del train, amb un model entrenat amb el train, seran tots òptims i totes les mètriques ens donaran 1.

Recordo que en el meu cas he optat per no fer partició de validation degut a l'escassetat de dades del nostre dataset.

Veiem com el test compta amb uns grans resultats entre 0.72 i 0,76 en totes les mètriques d'avaluació. Aquest bon rendiment ens demostra la robustesa del nostre model KNN, especialment tenint en compte la complexitat del conjunt de dades amb el qual hem treballat.

L'**accuracy** (que es mesura com el percentatge de prediccions correctes sobre el total de casos) té un valor de 0.753968.

La **precisió** (que es mesura com el percentatge de prediccions positives correctes sobre el total de prediccions positives) té un valor de 0.733552.

El **recall** (que es mesura com el percentatge de casos positius reals que han estat correctament identificats pel model) és de 0.753968.

L'**F1 Score** (que es mesura com la mitjana harmònica entre la precisió i el recall, proporcionant un balanç entre aquestes dues mètriques) és de 0.729078.

En conclusió, veiem per tant com el nostre model en KNN té bastant bons resultats en totes les mètriques.

5. Model Card

He aconseguit que em funcionés el ModelCard toolkit a Google Collab, utilitzant una mica les instruccions del laboratori 9 així m'ha quedat el Model Card:

Model Card for Cirrhosis Patient Survival Prediction Dataset

Model Details

Overview

Aquest model prediu la supervivència de pacients amb cirrosi hepàtica, donades 17 característiques clíniques.Aquest model està entrenat amb un algorisme de KNN. És un algorisme paràmetric i quees basa en la distancia manhattan als 19 veïns més propers per assignar mostres. Els hiperparàmetres que s'han hagut de controlar són el número de veïns, la mètrica i la distància.

Version

name: 1.0
date: 2023-12-28

Owners

Jaume Mora i Ladària (Alumne del Grau en Intel·ligència Artificial), jaume.mora.ladaria@estudiantat.upc.edu

References

- https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1

Considerations

Intended Users

- Professors i estudiants de IAA

Use Cases

- La intenció d'ús d'aquest model és per un treball en el qual he d'aprendre a utilitzari comparar diferents models, mitjançant una tasca d'aprenentatge supervisat en el qual es pretén predirla supervivència o no d'uns pacients amb cirrhosi hepàtica. Aquest model no té cap intenció en ser usat per arealitzar cap diagnostic real o negoci.

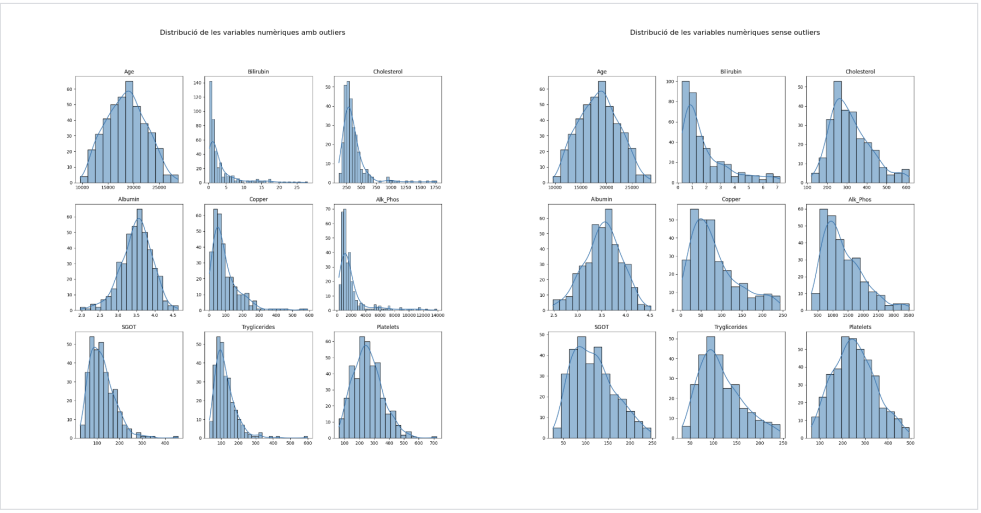
Limitations

- Les limitacions del KNN són diverses: sensibilitat a la mida del dataset i a la K, ineficiènica amb els datasets grans,sensibilitat a les variables irrellevants, necessitat d\'una normalització i/o estandardització de les dades i sensibilitat als outliers.

Ethical Considerations

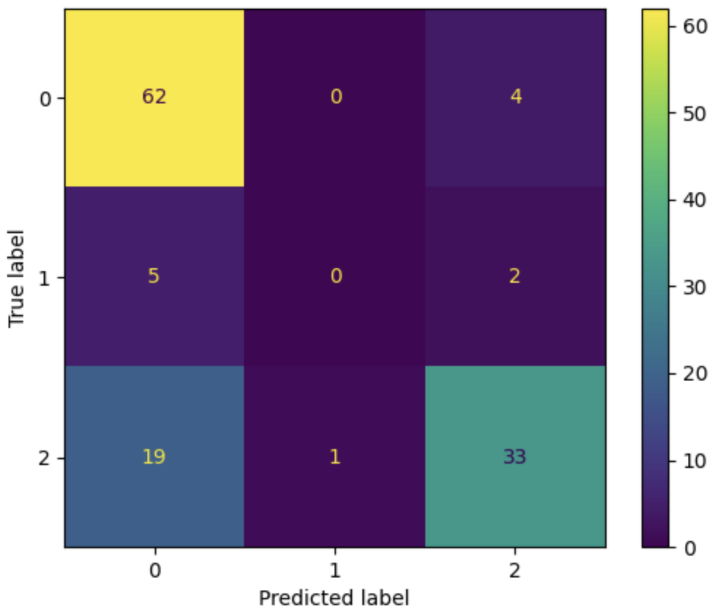
- Risk: Hem de complir les normes de privacitat en tot moment, estem tractant amb dades mèdiques. Mitigation Strategy: S'estan complint.

Datasets



Quantitative Analysis

Matriu de confusió

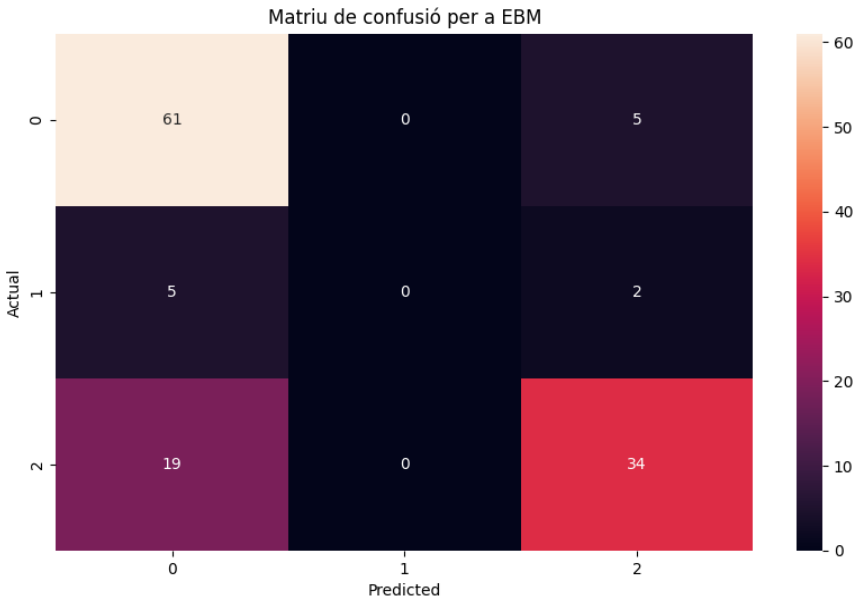


6. Bonus

6.1. Model Extra (EBM)

	Model	Accuracy	Precision	Recall	F1 Score
0	SVM	0.753968	0.780285	0.753968	0.727499

Resultats de l'EBM.



Matriu de confusió de l'EBM.

Veiem com el model EBM és lleugerament millor als models anteriors. De totes maneres, té un rendiment molt similar als altres. No prediu cap variable per a la classe 1 tampoc i encerta gairebé totes les variables de la classe 0. En aquest cas no he fet GridSearchCV perquè quan ho he intentat he vist que el temps d'execució era extremadament alt (anava pels 11 minuts i seguia executant).

7. Conclusions

En conclusió, estic bastant content amb el resultat del meu treball i crec que s'hi reflexen perfectament totes les hores dedicades. M'ha fet anar més enllà del que ja sabia sobre models, que tot i que ja havia fet alguna Hackathon i Datathon entrenant models, em pensava que en sabia molt més del que realment en sabia.

Crec que he processat prou bé totes les dades, potser m'ha faltat balancejar bé les dades, tot i que quan ho he intentat els resultats sempre han anat a pitjor. Per això no ho he acabat fent. De totes maneres, veig que els quatre models que he entrenat donen bastants bons resultats i això només m'encoratja a seguir aprenent en el camp de l'aprenentatge automàtic.

Jaume Mora i Ladària
Desembre del 2023