Universitat de les Illes Balears

TREBALL DE FI DE GRAU

ANÀLISI DE LA INCERTESA EN MODELS D'INTEL·LIGÈNCIA ARTIFICIAL APLICATS A DADES BIOMÈDIQUES

Francesc Llabrés Massanet

Grau de Matemàtiques

Escola Politècnica Superior

Any acadèmic 2024-25

ANÀLISI DE LA INCERTESA EN MODELS D'INTEL·LIGÈNCIA ARTIFICIAL APLICATS A DADES BIOMÈDIQUES

Francesc Llabrés Massanet

Treball de Fi de Grau

Escola Politècnica Superior

Universitat de les Illes Balears

Any acadèmic 2024-25

Paraules clau del treball: incertesa, conjunts creïbles, SMOTE, Gower

Tutors: Pedro Bibiloni Serrano i Arnau Mir Torres

SUMARI

Resum

En els models predictius aplicats a dades biomèdiques la incertesa pot tenir conseqüències crítiques. L'objectiu principal d'aquest treball serà el d'analitzar i caracteritzar aquesta incertesa en els nostres models, centrant-nos en la predicció de l'eficàcia d'un tractament per a l'hepatocarcinoma.

Proposam un mètode de quantificació de la incertesa basat en els conjunts creïbles. Analitzarem aquest mètode amb una base de dades senzilla i acadèmica, el *Wisconsin Breast Cancer Dataset*. A continuació, treballarem sobre una base de dades de l'Hospital Universitari de Son Espases, de 144 pacients i 118 variables, que presentarà una major complexitat.

Per a fer prediccions, s'han implementat tècniques com el *bootstrap*, LASSO, LOOCV i SMOTE, i mètodes de classificació com regressió logística i *random forest*. A més, s'ha desenvolupat un mètode d'imputació de valors buits basat en la distància de Gower i els *K-nearest neighbors* i un mètode per a extreure informació sobre quines variables ens aporten més informació mitjançant *bootstrap* i LASSO.

Com a contribucions, el treball ofereix un mètode efectiu per a quantificar la incertesa, una estratègia d'imputació de dades robusta i informació sobre les variables més influents en la resposta al tractament. Hem pogut veure, per exemple, que aspectes en principi irrellevants com la localització tumoral afecten a l'efectivitat del tractament. A més, malgrat que el nostre model final obtingui una alta incertesa degut a la complexitat del problema, hem aconseguit que la capacitat predictiva de l'estudi sigui millor que la que teníem fins al moment, amb una presició del 66%.

CAPÍTOL 1

INTRODUCCIÓ

L'estadística i l'anàlisi de dades juguen un paper fonamental en la ciència moderna, especialment en àmbits on la incertesa és inherent, com la medicina. La presa de decisions en aquest context es basa sovint en dades incompletes o incertes, fet que pot tenir conseqüències crítiques per als pacients. Comptar amb models predictius fiables no és suficient; també és essencial quantificar la incertesa associada a aquestes prediccions. Per a fer-ho, analitzarem i quantificarem la incertesa mitjançant conjunts creïbles, que ens permetran obtenir més informació sobre la quantitat i tipus d'incertesa extreta i millorar la interpretabilitat dels resultats. En aquest treball, ens centrem en l'anàlisi de la incertesa en models d'intel·ligència artificial aplicats a la predicció de l'eficàcia d'un tractament específic per a l'hepatocarcinoma, anomenat quimioembolització transarterial, o TACE.

L'hepatocarcinoma és el tipus més comú de càncer de fetge i una de les principals causes de mort per càncer a nivell mundial. Sovint es diagnostica en estadis avançats, fet que limita les opcions terapèutiques i redueix les possibilitats de supervivència. No obstant això, la resposta als tractaments varia significativament entre pacients, fet que fa essencial el desenvolupament d'eines predictives que permetin personalitzar les decisions terapèutiques.

L'objectiu principal del treball és caracteritzar adequadament la incertesa associada a les prediccions que faran els nostres models, un aspecte clau en un context on els tractaments poden ser invasius i comportar riscos significatius. A més, estudiarem diferents maneres d'imputar valors buits i identificarem les variables més rellevants. Per dur a terme aquesta anàlisi, es treballarà amb un conjunt de dades proporcionat per l'Hospital Universitari de Son Espases (HUSE), que presenta una quantitat considerable de valors buits i un gran nombre de variables, moltes de les quals podrien no tenir una relació clara amb l'eficàcia del tractament. En aquest context, s'utilitzaran diverses tècniques d'imputació per tractar els valors absents i es provaran diferents mètodes d'aprenentatge automàtic, com la regressió logística amb LASSO i els Random Forest. A

més, s'aplicaran tècniques com el bootstrap, la validació creuada LOOCV i l'algorisme SMOTE per millorar la robustesa dels resultats.

Els principals reptes d'aquest estudi són la gestió de la informació incompleta i la identificació de les variables rellevants per a la predicció. Malgrat l'alta complexitat de la base de dades a treballar, esperem obtenir diverses aportacions significatives, com un mètode per quantificar i interpretar millor la incertesa en aquests models, estratègies eficients d'imputació de dades i una millor comprensió de les variables més rellevants en la resposta al nostre tractament.

1.1 Estructura del treball

Per millorar la comprensió, després del capítol d'introducció presentarem un resum dels conceptes preliminars essencials per al desenvolupament del projecte. Aquest inclourà les eines estadístiques emprades, els principis teòrics sobre la incertesa i la seva quantificació, així com una descripció de les tres bases de dades utilitzades.

Després, explicarem tots els experiments realitzats i els seus resultats, dividits en tres blocs, un per a cada base de dades. Finalment, resumirem els nostres resultats més significatius, les conclusions que en podem extreure, i les possibles línies de recerca futures.

1.2 Repositori

El codi usat per a fer el preprocessament de dades, els models i el seu posterior anàlisi, el podem trobar en el següent repositori de GitHub:

`https://github.com/XescLlabres/tfgFLM.git`

Per a resumir la nostra contribució en forma de codi, i per tal d'esquematitzar i referenciar tots els experiments realitzats, tenim la següent taula:

Taula 1.1: Taula d'experiments realitzats.

Experiment	**Secció**	**Nom del fitxer**	**Descripció**
Model 1: Regressió logística	3.1.1	TaulaLOOCV	Preprocessament i entrenament del model 1.
Model 1: Avaluació	3.1.2	AvaluacioModel	Mètriques de rendiment del model 1.
Model 1: Avaluació incertesa	3.1.3	TPUncertainty	Anàlisi sobre la incertesa obtinguda.
Variació intervals creïbles	3.1.4	INT	Modificar grau de confiança i veure com varia la incertesa.
Model 1: Introducció de renou	3.1.5	RENOU	Introducció renou i anàlisi de la incertesa.
Preprocessament	3.2.1	dadeshuse	Preprocessament de nova base de dades per pacient.
Comprovació de mètodes d'imputació	3.2.1	MitjanaModa, KNNImputer, Gower	Comprovació i comparació dels 3 mètodes d'imputació de valors buits.
Model 2: Regressió logística i LASSO	3.2.3	Gower3	Entrenament del model amb regressió logística i LASSO.
Model 2: Introducció de SMOTE	3.2.3	GowerSMOTE	Mateix model amb tècnica de SMOTE.
Model 3: Random forest	3.2.4	RandomForest, RFtrainig	Entrenament del model amb random forest.
Preprocessament	3.3.1	preprocestumors	Preprocessament de la base de dades per massa tumoral.
Model 4: Random forest	3.3.2	rforestMASSTUMORALS	Entrenament del model amb random forest.
Model 5: Regressió logística	3.3.3	MASSTUMORALS	Entrenament del model amb regressió logística.
Model 5: Anàlisi incertesa	3.3.4	AnalisiMASSTUMORALS	Avaluació del model i de la incertesa obtinguda.

CAPÍTOL 2

CONCEPTES PREVIS

2.1 Eines estadístiques

L'objectiu d'aquesta secció és explicar de manera resumida totes les eines estadístiques usades dins de l'experimentació, per tal de facilitar més endavant la comprensió dels models i conclusions extretes.

La informació tècnica d'aquesta secció la podem trobar en el llibre *An Introduction to Statistical Learning with Applications in Python* [1], àmpliament conegut i usat dins l'àmbit estadístic i de la ciència de dades. Per a la tècnica de SMOTE i la distància de Gower, que surten de l'abast del llibre, hem trobat tota la informació detallada als articles *SMOTE: Synthetic Minority Over-sampling Technique* [2] i *A general coeficient of similarity and some of its properties* [3], respectivament.

2.1.1 Bootstrap

El *bootstrap* és un mètode basat en el mostreig amb reemplaçament que es fa servir, per exemple, per a fer estimacions d'hiperparàmetres o per a calcular intervals de confiança d'un cert estimador. És especialment eficient quan no disposam d'un gran nombre d'observacions.

L'objectiu és dur a terme moltes mostres amb reemplaçament d'un mateix conjunt de dades, i calcular l'estadístic d'interés per a cada una d'elles i així poder obtenir una distribució d'aquest. Una de les aplicacions més importants del *bootstrap* és la construcció d'intervals de confiança per a estimacions. Es realitza de la següent manera:

1. Es parteix d'una mostra original de dades $\{x_1, x_2, \ldots, x_n\}$ de mida n.
2. Es generen múltiples mostres bootstrap (generalment 1.000 o més), cadascuna de mida n, seleccionant aleatòriament observacions de la mostra original amb reemplaçament.

3. Per a cada mostra bootstrap, es calcula l'estadístic d'interès.

4. Es defineixen els límits de l'interval de confiança calculant els percentils corresponents.

Una limitació d'aquest mètode és que és computacionalment costós, però a l'estar treballant amb un nombre baix de casos en totes les bases de dades, no ens suposarà un problema.

2.1.2 Cross Validation

La validació creuada (*cross validation*) és una tècnica d'avaluació i validació de models que permet obtenir una estimació fiable del seu rendiment i saber si aquest s'adapta bé a les dades.

El conjunt de dades es divideix en diverses parts (anomenades *folds*) i el model s'entrena i es valida repetidament en diferents subconjunts de les dades. L'algorisme és el següent:

1. Es divideix el conjunt de dades en K parts iguals, on $K \in \{2,3,...,N\}$ i N és la grandària de la taula de dades.

2. Es selecciona un dels folds com a conjunt de validació, mentre que els $K-1$ restants s'utilitzen per a entrenar el model.

3. Aquest procés es repeteix K vegades, utilitzant cada fold com a conjunt de validació una vegada.

4. Es calcula la mètrica de rendiment desitjada en cada iteració i finalment es fa la mitjana dels resultats per obtenir una estimació final del rendiment del model.

El *Leave-One-Out Cross-Validation* (LOOCV) és una tècnica que pertany a la família dels mètodes de *cross validation*. Funciona deixant una única observació fora del conjunt d'entrenament i utilitzant-la com a conjunt de validació, és a dir, el cas concret on $K = N$. És especialment útil per avaluar el rendiment d'un model predictiu quan el conjunt de dades és relativament petit, com el nostre.

El procés del LOOCV es pot resumir en els següents passos:

1. Es parteix d'un conjunt de dades amb n observacions: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

2. Per a cada observació i:

   - Es deixa fora la i-èsima observació del conjunt d'entrenament.
   - Es construeix el model utilitzant les $n-1$ observacions restants.
   - Es fa una predicció de la i-èssima observació fent servir el model estadístic.

Una altra vegada, és un mètode computacionalment costós, però no ens suposarà un problema al tractar amb un nombre baix d'observacions.

2.1.3 Mètodes de classificació

En aquesta subsecció, veurem els mètodes que hem fet servir per a classificar.

Regressió Logística

La regressió logística és una tècnica que serveix per modelar problemes de classificació binària, on la variable resposta Y pren només dos valors possibles. Aquesta tècnica estima la probabilitat que un esdeveniment pertanyi a una de les dues categories, basant-se en un conjunt de predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

El model es basa en la funció logística, que transforma una combinació lineal dels predictors en la probabilitat de $Y = 1$ donat un valor concret dels predictors. Formalment, l'expressió de la probabilitat que un conjunt de predictors $\mathbf{X}$ pertanyin a la classe 1 és:

$$P(Y = 1 \mid \mathbf{X}) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}, \tag{2.1}$$

on β_0 és l'ordenada a l'origen i $\beta_1, \ldots, \beta_p$ són els coeficients associats als predictors.

Aquests paràmetres $\beta_0, \beta_1, \ldots, \beta_p$ es calculen utilitzant el mètode de màxima versemblança. L'objectiu és trobar els valors dels paràmetres que maximitzin la probabilitat de les observacions donades:

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} P(Y = 1 \mid \mathbf{X}_i)^{y_i} \cdot P(Y = 0 | \mathbf{X}_i)^{1 - y_i}, \tag{2.2}$$

on y_i representa l'etiqueta observada de la mostra i-èssima, i $P(Y = 1 \mid \mathbf{X}_i)$ és la probabilitat estimada que $Y = 1$ per aquesta mostra.

La regressió logística presenta una gran flexibilitat i interpretabilitat, que ens serà de gran utilitat en els nostres models.

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO és una tècnica utilitzada en regressió per introduir penalitzacions als coeficients del model, amb l'objectiu de millorar-ne la interpretabilitat i reduir el risc de sobreajustament. És molt útil en models on hi hagi gran quantitat de variables predictores, com és el nostre cas, ja que els coeficients associats a variables no rellevants són reduïts exactament a zero, realitzant així una selecció automàtica de variables i simplificant el model.

LASSO és àmpliament coneguda i utilitzada en models lineals. Malgrat això, la podem utilitzar en el context de la regressió logística en forma de penalització sobre la suma dels valors absoluts dels coeficients, β_j, en el problema d'optimització, eliminant així les variables no rellevants del model. La funció objectiu per a la regressió logística amb penalització LASSO es basa en la màxima versemblança regularitzada, seguint la notació anterior tenim:

$$\min_{\beta_0, \beta} \left[-\ell(\beta_0, \beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.3}$$

on $\ell(\beta_0, \beta)$ és la funció de log-versemblança del model logístic:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \left[y_i \log P(Y = 1 \mid \mathbf{X}_i) + (1 - y_i) \log(1 - P(Y = 1 \mid \mathbf{X}_i)) \right], \tag{2.4}$$

on recordem que $P(Y = 1 \mid \mathbf{X}_i)$ és la probabilitat predita que $Y = 1$, definida com:

$$P(Y = 1 \mid \mathbf{X}_i) = \frac{e^{\beta_0 + \mathbf{X}_i^T \beta}}{1 + e^{\beta_0 + \mathbf{X}_i^T \beta}}, \tag{2.5}$$

on $\mathbf{X}_i$ és el vector de predictors per a l'observació i-èssima, β_0 és l'ordenada a l'origen i $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ són els coeficients del model. El terme $\lambda \geq 0$ és el paràmetre de regularització que controla la força de la penalització.

Aquest paràmetre λ és fonamental per equilibrar el compromís entre el biaix i la variància. Un valor elevat de λ augmenta la penalització, eliminant més variables però incrementant el biaix. Per contra, un valor petit de λ permet que més variables tinguin coeficients diferents de zero, però pot provocar sobreajustament. Nosaltres calcularem aquest paràmetre mitjançant un *grid search*, explicat a la secció 2.1.4, per tal d'optimitzar la precisió del model.

En definitiva, usarem LASSO principalment perquè ens trobarem davant models amb un gran nombre de predictors, que és un problema típic quan es tracten dades mèdiques. A més, aquesta selecció ens permetrà no només millorar la interpretabilitat i reduïr el risc de sobreajustament, sinó també saber quines d'aquestes variables són més importants dins els nostres models, que és una de les preguntes a les que intentam donar resposta.

Arbres de Decisió

Un arbre de decisió és un model predictiu que segmenta l'espai de característiques en regions homogènies segons una sèrie de divisions successives basades en els valors dels predictors. Per construir un arbre de decisió per a classificació, es duen a terme els següents passos:

1. Es selecciona la millor variable i punt de tall per dividir les dades en dues subregions, basant-se en una mesura de puresa com l'índex de Gini (2.6) o l'entropia (2.7).

2. Es repeteix el procés recursivament en cada subregió fins que es compleixi un criteri d'aturada, com ara una profunditat màxima o un nombre mínim d'observacions per fulla.

3. Les fulles de l'arbre contenen la predicció final, que correspon a la classe majoritària de les dades en aquella fulla, la moda.

Per determinar quina variable i punt de tall utilitzar en cada divisió, s'empren mesures com:

- **Índex de Gini:** Mesura la probabilitat que una observació seleccionada a l'atzar sigui classificada incorrectament si es fa una predicció basada en la distribució de classes de la regió. Es defineix com:

$$G = 1 - \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.6}$$

  on $\hat{p}_{mk}$ és la proporció d'observacions en la regió m-èssima de la classe k-èssima.

- **Entropia:** Mesura el desordre de la distribució de classes en una regió. Es calcula com:

$$H = -\sum_{k=1}^{K} p_k \log(p_k), \tag{2.7}$$

  on p_k té el mateix significat que en l'índex de Gini.

Random Forest

Aquesta explicació del arbres de decisió ens és d'utilitat per a veure com funciona un *random forest*, que és un mètode d'aprenentatge automàtic utilitzat tant per a problemes de regressió com, en el nostre cas, de classificació. Es basa en la combinació d'un gran nombre d'arbres de decisió construïts de manera independent per millorar l'exactitud de les prediccions i reduir el risc de sobreajustament.

El *random forest* es construeix mitjançant el següent procediment:

1. **Generació de mostres bootstrap:** Es generen B mostres bootstrap a partir del conjunt de dades original. Cada mostra és generada seleccionant aleatòriament observacions amb reemplaçament, fet que garanteix que cada mostra tingui una combinació única d'observacions.

2. **Construcció d'arbres de decisió:** Per a cada mostra bootstrap, es construeix un arbre de decisió utilitzant un subconjunt aleatori de variables en cada bifurcació. Aquesta selecció aleatòria de variables ajuda a reduir la correlació entre els arbres i millora la diversitat del conjunt.

3. **Predicció:** Cada arbre realitza una predicció per al nou exemple, i el resultat final es decideix per vot majoritari.

Un dels punts clau del *random forest* és la selecció d'un subconjunt aleatori de variables a cada bifurcació. Aquest enfocament redueix la probabilitat que arbres diferents facin les mateixes divisions, incrementant la diversitat entre els arbres i reduint el risc de sobreajustament. El nombre de variables seleccionades m sovint es fixa com $m = \sqrt{p}$ per a classificació, on p és el nombre total de predictors.

2.1.4 Recerca d'hiperparàmetres

La recerca d'hiperparàmetres és el procés d'ajustar els paràmetres que s'han de definir abans de l'entrenament del model. El *grid search* és una tècnica que permet provar diferents combinacions d'hiperparàmetres per trobar la millor configuració segons una

mètrica d'avaluació.

Funciona de la següent manera:

1. Es defineix una graella de possibles hiperparàmetres.
2. S'entrena el model amb validació creuada, amb totes les combinacions possibles.
3. Es mostra la selecció d'hiperparàmetres amb millors resultats.

Per tal d'optimitzar els nostres models de regressió logística amb LASSO i de random forest, realitzam aquesta recerca per a trobar els hiperparàmetres que millor s'ajustin:

- En el cas de la regressió logística amb LASSO, ho feim per a trobar el grau de regularització òptim en fer la selecció de variables, és a dir, el millor λ.
- En el random forest, en canvi, hi trobam més hiperparàmetres per definir: el nombre d'arbres, la profunditat màxima d'aquests, el mínim de mostres per dividir un node o el mínim de mostres per conservar una fulla.

2.1.5 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (Synthetic Minority Over-sampling Technique) és una tècnica d'augmentació de dades dissenyada per resoldre el problema de desequilibri de classes en problemes de classificació. Quan es té una classe majoritària molt més nombrosa que la classe minoritària, els models d'aprenentatge automàtic poden tenir dificultats per identificar patrons rellevants de la classe minoritària. SMOTE genera instàncies sintètiques per a la classe minoritària mitjançant la interpolació entre mostres de la classe minoritària existent, en lloc de replicar directament les existents, millorant així el rendiment dels classificadors.

Es generen les mostres sintètiques de la següent manera:

1. **Identificació de veïns propers:** Per cada mostra de la classe minoritària, es calculen els k veïns (5 per defecte) més propers fent servir una distància predeterminada, la euclídea en el nostre cas.
2. **Generació de mostres sintètiques:**
   a) Es selecciona aleatòriament un dels k veïns, el $x_{\text{veïnat}}$.
   b) Es crea una nova instància sintètica x_{new} a partir d'una de ja existent, x_{original}, utilitzant la fórmula següent:

$$x_{\text{nou}} = x_{\text{original}} + \lambda(x_{\text{veïnat}} - x_{\text{original}})$$

   on λ és un valor aleatori entre 0 i 1 que determina on es troba el nou punt sintètic al segment entre $x_{\text{veïnat}}$ i x_{original}. És a dir, si $\lambda = 0$ el punt sintètic serà exactament l'original, si $\lambda = 1$ serà exactament el veïnat i si $\lambda \in (0,1)$ no coincidirà exactament amb cap dels dos.

3. Si es vol augmentar la mida de la classe minoritària en un N%, es repeteix el procés generant el nombre corresponent de mostres sintètiques. Per exemple, si es vol un 200% d'augment, es generen dues noves mostres per cada mostra original.

Aquesta tècnica ajuda a equilibrar les classes, millorant així la capacitat del model per aprendre de la classe minoritària.

2.1.6 Distància de Gower

La distància de Gower és una mètrica de similaritat generalitzada que permet comparar individus en funció de variables de diferents tipus. Ens ha estat de gran ajuda ja que a les nostres dues bases de dades mèdiques, trobam tant dades quantitatives com categòriques.
Es defineix de la següent manera: per a dos individus i i j, la similaritat de Gower S_{ij} es calcula com la mitjana ponderada de les similaritats per cada variable k:

$$S_{ij} = \frac{\sum_{k=1}^{p} s_{ijk} \cdot w_k}{\sum_{k=1}^{p} w_k},$$

on:

- s_{ijk} és la similaritat entre els individus i i j per la variable k.
- w_k és el pes assignat a la variable k (en el nostre cas $w_k = 1$).
- p és el nombre total de variables.

La distància de Gower s'obté com:

$$D_{ij} = 1 - S_{ij}.$$

On cada tipus de variable té la seva pròpia manera de calcular la similaritat:

- **Variables quantitavtives:**
$$s_{ij}^k = 1 - \frac{|x_i^k - x_j^k|}{R^k},$$
on R^k és el rang, definit com $R^k = \max(x^k) - \min(x^k)$.
- **Variables categòriques:**
$$s_{ij}^k = \begin{cases} 1, & \text{si } x_i^k = x_j^k, \\ 0, & \text{si } x_i^k \neq x_j^k. \end{cases}$$

Com a aspectes negatius, trobam que la similaritat en variables categòriques tindrà un pes major que la de les quantitatives, ja que aquesta segona tindrà valors entre 0 i 1, mentre que la primera tindrà un dels dos valors extrems.

Així i tot, aquesta distància ens soluciona el problema de la varietat de dades, que ens limitava molt, per exemple, a l'hora de trobar veïns per imputar valors buits.

2.1.7 Mesura de la qualitat d'un model

Per avaluar el rendiment d'un model de classificació, es poden utilitzar diverses mètriques basades en la matriu de confusió, que és la taula que resumeix el rendiment d'un model de classificació segons els encerts i els errors:

Taula 2.1: Tipus d'errors de classificació.

	Predicció Negativa	**Predicció Positiva**
Classe Negativa	True Negatives (TN)	False Positives (FP)
Classe Positiva	False Negatives (FN)	True Positives (TP)

A partir d'aquesta taula, trobam diverses mètriques per a poder avaluar el nostre model. Usarem les següents:

Taula 2.2: Mètriques de rendiment.

Mètrica	**Fórmula**
Precisió	$\frac{TP}{TP+FP}$
Recall (Sensibilitat)	$\frac{TP}{TP+FN}$
F1-Score	$2 \cdot \frac{\text{Precisió} \cdot \text{Recall}}{\text{Precisió} + \text{Recall}}$
Exactitud	$\frac{TP+TN}{TP+TN+FP+FN}$

La corba ROC (*Receiver Operating Characteristic*) representa la relació entre el *True Positive Rate* (Recall) i el *False Positive Rate* (FPR):

$$\text{FPR} = \frac{FP}{FP+TN}.$$

L'àrea sota la corba ROC (AUC) proporciona una mesura global del rendiment del model: valors propers a 1 indiquen un model excel·lent, mentre que valors propers a 0.5 indiquen un model aleatori.

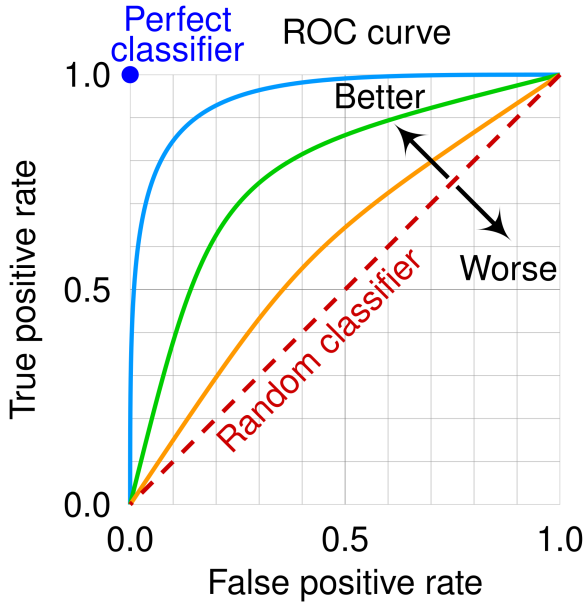


Figura 2.1: Corba ROC, imatge de Wikipèdia [4].

2.2 Incertesa

Dins l'àmbit de l'anàlisi de dades, entenem per incertesa com la manca de seguretat o de precisió sobre un model predictiu. És una part fonamental per a la presa de decisions i, al tractar amb fenòmens de la vida real, inevitable. Dins el camp de la medicina, comprendre, quantificar i gestionar aquesta incertesa és essencial, ja que els errors en aquest context poden tenir conseqüències greus per als pacients.

Aquesta incertesa ens arriba de molts llocs diferents. Les dades limitades, amb renou o incompletes dificulten qualsevol tipus de predicció, com la variabilitat en els processos biològics, on cada pacient és un món, o simplement la complexitat del fenòmen clínic a estudiar. Segons el treball de Fox i Ülkünen [5], podem dividir la incertesa en dos grans blocs: la incertesa aleatòria i la incertesa epistèmica.

2.2.1 Incertesa aleatòria

La incertesa aleatòria es refereix a la variabilitat inherent i imprevisible d'un fenòmen o procés. És irreductible i forma part intrínseca del sistema que s'està estudiant. És la incertesa restant quan estam tractant d'un model perfecte i amb informació completa. Prové de distints punts, com la variabilitat intrínseca d'un fenomen aleatòri (com el llançament d'una moneda), la intervenció de variables externes no mesurables, o el renou inevitable que tenen segons quins processos de recollida de dades. Per exemple, per tornar al món de la medicina, el temps de resposta d'un pacient a un tractament pot variar fins i tot si les condicions inicials són idèntiques, degut a diferències biològiques no mesurades.

No estarà a les nostres mans la possibilitat de reduïr-la, i ens indicarà les limitacions intrínseques del que estam predint.

2.2.2 Incertesa epistèmica

La incertesa epistèmica es refereix a la manca de coneixement o informació sobre el fenòmen que s'està modelant. A diferència de la aleatòria, és reduïble i ens donarà una percepció de com de bo és el nostre model. També prové de distintes fonts, com tenir dades insuficients, incompletes o errònies, que el model sigui massa simple o utilitzi assumpcions incorrectes, o que el model no generalitzi bé perquè està analitzant només un conjunt específic de la població estudiada.

El fet de que es pugui reduïr, implica que la predicció pot millorar amb més i millors dades, adaptant el model a les característiques de l'estudi, i serà un bon indicador de si un model en concret és bo o no.

2.2.3 Conjunts creïbles

Per tal de poder quantificar aquesta incertesa i poder seguir amb el nostre estudi, necessitam introduir un nou concepte, els conjunts creïbles, o *credal sets*. Com nosaltres treballam dins d'un espai de probabilitats unidimensional, ens centrarem únicament en els **intervals creïbles**, però aquest concepte es pot extendre a altres dimensions, vegeu els articles de Paul Hofman *et al.* [6] i Eyke Hüllemeier *et al.* [7] per a més informació.

Quan treballem amb models predictius que assignen probabilitats a un resultat, sovint no podem conèixer aquesta probabilitat amb exactitud. En aquests casos, en lloc de treballar amb un únic valor, podem utilitzar un interval creïble, que és el conjunt de valors que considerem compatibles amb la informació disponible.

En aquest treball, calcularem els intervals creïbles seguint un procediment similar al que s'utilitza per trobar intervals de confiança en estadística, encara que aquests s'utilitzin per a estimar paràmetres.

El mètode consisteix en:

1. **Simular moltes vegades el càlcul de la probabilitat**: Generem múltiples valors per p, basant-nos en la distribució de probabilitat del nostre model.

2. **Construir un interval que cobreixi la major part d'aquests valors**: Ordenem els valors simulats i seleccionem l'interval que conté el 95% (o el nivell de confiança desitjat) d'aquests valors a partir dels percentils dels valors trobats.

Aquest procés ens permet obtenir intervals creïbles dins de l'espai de probabilitats [0,1], que reflecteixen tant la informació de les dades com la incertesa del model. Vegem-ho amb un exemple:

Si un model predictiu estima que la probabilitat d'un esdeveniment és $p = 0.7$, podríem assumir que aquest valor és precís. Però en realitat, si el model té incertesa

(perquè les dades són limitades o el fenomen és complex), és més informatiu expressar la probabilitat com un interval, per exemple:

$$p \in [0.6, 0.8]$$

El mateix esdeveniment però amb unes altres dades o un altre model ens diu que la probabilitat de l'esdeveniment és també $p = 0.7$ però ens retorna un interval creïble de la següent forma:

$$p \in [0.69, 0.71]$$

Si treballàssim tan sols amb probabilitats, ambdós models ens dirien que la probabilitat d'aquest determinat esdeveniment és del 70%, però al treballar amb intervals creïbles, tenim una visió molt més general i precisa d'on recau aquesta probabilitat, ja que en el primer cas el rang és major i hi trobam molta més incertesa, mentre que amb el segon model podrem afirmar amb bastanta seguretat que la probabilitat real serà pròxima al 70%.

Per tant, els intervals creïbles ens ajudaran a expressar la incertesa sobre una probabilitat d'una forma molt més informativa que amb un valor únic, i veurem a continuació que la podrem quantificar molt fàcilment.

2.2.4 Quantificació de la incertesa

Ens trobam ara amb l'objectiu no només de poder quantificar i representar la incertesa dels que seran els nostres models, sinó també de poder representar de manera numèrica quina quantitat d'aquesta és incertesa epistèmica i quina quantitat és aleatòria.

Quan treballam amb rangs de probabilitats en una dimensió, on els únics valors possibles siguin 0 o 1, i.e, una classificació binària, cada predicció es representa amb un interval de la forma $[a, b]$, on a és la probabilitat mínima assignada a l'esdeveniment (en el nostre cas, dins el nostre interval creïble), i b la probabilitat màxima. Per quantificar la incertesa, no ens interessa només que aquest interval sigui petit, sino que sigui pròxim a un dels dos extrems, ja que com més pròxim estigui l'interval a 0.5, més incertesa tindrà la nostra predicció.

Hullemeier *et. al* [7] proposen una quantificació de la incertesa total de la forma

$$TP(a, b) = \min(1 - a, b).$$

Aquesta mètrica arriba a valors màxims quan l'interval abarca tot el rang de probabilitats $[0, 1]$ i decreix a mesura que l'interval es fa petit, representant així la incertesa epistèmica de la predicció. Quan feim desaparèixer aquesta incertesa epistèmica, i.e $a = b$, la mètrica té el seu valor màxim dins aquests casos quan $a = b = 1/2$, i es va reduïnt a 0 a mesura que s'atraca a un dels dos extrems.

Podrem descomposar-la seguint aquest raonament:

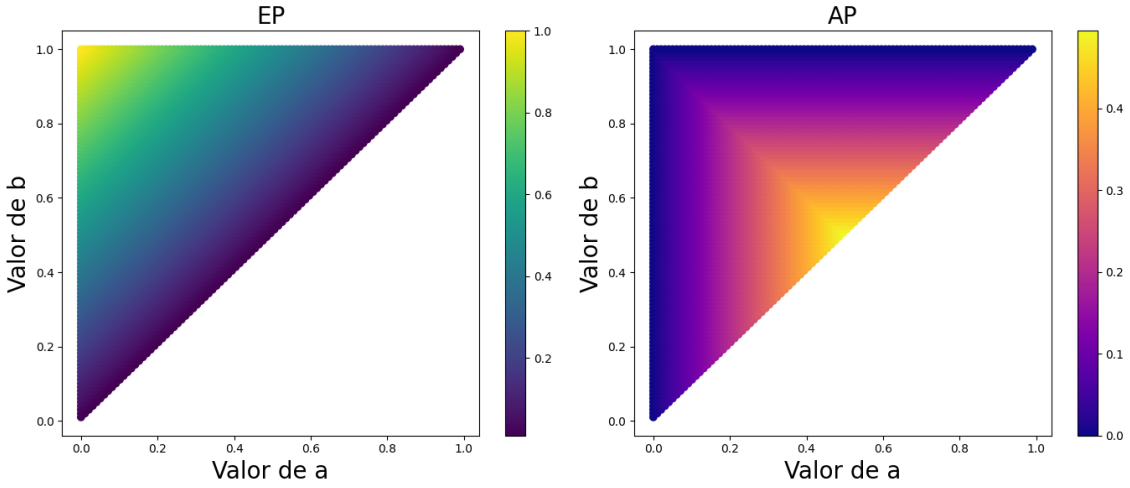
- La incertesa epistèmica (EP), com a l'amplada de l'interval, $b - a$.

- La incertesa aleatòria (AP), com a la proximitat de la predicció a un dels dos casos, i.e $\min(a, 1-b)$

Arribant doncs a

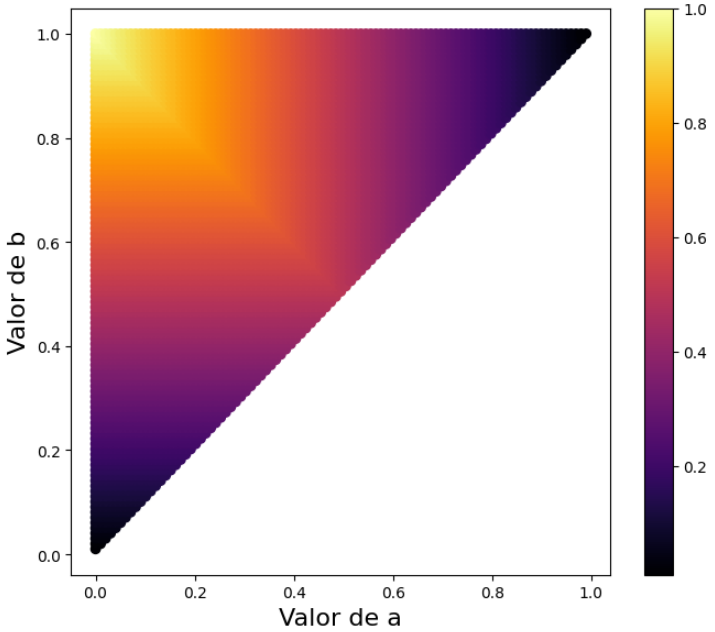
$$TP(a,b) = \min(1-a,b) = \min(a,1-b) + (b-a) = AP(a,b) + EP(a,b).$$

Figura 2.2: Mapes de calor de $EP(a,b) = b-a$ i $AP(a,b) = \min(a,1-b)$.



Podem veure en el primer gràfic, que la incertesa epistèmica augmenta a mesura que el valor de a i el valor de b s'allunyen. En el segon, la incertesa aleatòria augmenta com més pròxim a 0.5 està l'interval.

Figura 2.3: Mapa de calor de $TP(a,b) = \min(1-a,b)$.



En el mapa de calor 2.2.4, podem veure que obtindrem menys incertesa a mesura que el valor de a i de b disminueixi, i.e. l'interval creïble sigui petit, i a mesura que ens allunyem del 0.5.

Aquesta quantificació ens aporta una forma senzilla i clara de representar la incertesa, i permet identificar amb precisió la proporció d'incertesa atribuïble a la manca de coneixement (epistèmica) i a la variabilitat inherent (aleatòria).

Tot i això, el mètode assumeix que la incertesa epistèmica i aleatòria són mesurables de manera independent, i en casos complexos, aquestes dues formes d'incertesa poden no ser fàcilment separables.

2.3 Bases de dades

Durant el treball de recerca hem fet feina amb tres bases de dades. Com que les bases de dades reals a tractar ens presenten molts de problemes prèvis, i per tal de comprovar que tant el nostre model predictiu com la nostra forma de quantificar la incertesa sigui bona, hem realitzat una experimentació prèvia amb una base de dades acadèmica i senzilla, el *Wisconsin Breast Cancer*, una taula de dades biomèdiques, on cada fila representa un tumor i voldrem predir la presència o absència de malignitat.

Un cop fet aquesta investigació, ens hem centrat en una segona base de dades, proporcionada per l'Hospital Universitari de Son Espases. Aquesta base de dades representa tota la informació mèdica al nostre abast sobre pacients que pateixen hepatocarcinoma i que han estat sotmesos a un TACE. El nostre objectiu serà crear un model que pugui predir si aquest tractament serà o no eficient per a un determinat pacient.

Més endavant, adaptarem aquesta mateixa base de dades, per dos motius. El primer és que desde l'hospital ens proporcionen més dades, el que redueix el nombre de valors buits i millora el model. El segon motiu, i el més determinant, és que canviam l'enfoc de l'estudi. A partir d'aquesta base de dades, no intentam predir l'efectivitat del tractament per pacient, sino per massa tumoral. Ens adonam que diferents masses tumorals d'un mateix pacient responen diferent a l'estudi, i la variable predictora serà *Viable*, que valdrà 0 si el tumor s'ha reduit al 100% i valdrà 1 en cas contrari.

2.3.1 Wisconsin Breast Cancer

Es tracta d'una base de dades de 569 tumors, cadascú respresentat per una fila, i 32 columnes, on una vegada apartades la variable predictora i la columna amb el nombre d'identificació de cada pacient, el que obtenim són 30 columnes numèriques que representen distints aspectes mèdics sobre el càncer de mama a estudiar. La variable predictora es diu *diagnosis*, que representarem amb un 0 quan el tumor sigui beninge, i amb un 1 quan sigui maligne.

Normalment, el diagnosis de càncer de mama es fa mitjançant una biòpsia completa, però amb aquestes dades, recollides totes per imatge, podrem veure que amb aquest mètode molt menys invasiu, arribam a prediccions molt bones.

Tenim 30 columnes i 10 components del tumor a estudiar. Per a cada component, tenim tres valors guardats: la seva mitjana, la desviació típica i el *worst*, que és la mitjana dels tres valors més grans, arribant així a les 30 variables. Les components del tumor són les següents:

1. *Radi*: Es calcula la mitjana de distàncies del centre del nucli fins al seu contorn.
2. *Textura*: Representa la variació en la intensitat dels píxels dins del nucli.
3. *Perímetre*: Mesura la longitud del contorn del nucli.
4. *Àrea*: Calcula la mida total del nucli en píxels quadrats.

5. *Suavitat*: Mesura la regularitat dels contorns del nucli, es calcula com a la variació local de la longitud de les distàncies radials.
6. *Compacitat*: Relació entre l'àrea i el perímetre: $\frac{perimetre^2}{area-1.0}$
7. *Concavitat*: Mesura el grau de concavitat de les vores del nucli.
8. *Punts de concavitat*: Calcula el nombre de punts còncavs en el contorn del nucli.
9. *Simetria*: Compara la simetria entre les dues parts del nucli.
10. *Dimensió fractal*: Mesura la complexitat del contorn del nucli, es calcula com la relació entre el perímetre i l'àrea, però amb un anàlisi fractal.

Per a més informació sobre la recollida de dades i les variables d'aquesta taula de dades, es pot trobar en l'article de W. Nick Street *et al.* [8].

Aquesta base de dades no conté valors buits, és bastant extensa i ens conduirà a bons resultats i a incerteses baixes, el que ens servirà d'ajuda per a estudiar el nostre model i analitzar les nostres mètriques per a l'incertesa.

2.3.2 Base de dades per pacients

És molt més complexa que l'anterior. Consta de 144 pacients i 118 variables a considerar. La variable predictora és *RECIST* que correspont a la resposta del pacient al tractament, serà 1 si el tractament funciona (resposta total o millora), i 0 si no ho fa (malaltia estable o empitjorament).

Per tal d'explicar de manera resumida les variables, les dividirem en blocs:

- **Dades demogràfiques**: com podrien ser l'edat, el dia del tractament o el sexe.
- **Informació clínica**: com l'altura o el pes, els antecedents patològics (obesitat, diabetes, alcoholisme...), historial de malalties del fetge, fibrosis, o diversos *scores* sobre el tumor.
- **Informació de laboratori**: dades hematòlogues (de la sang), de coagulació, bioquímiques i marcadors serològics (substàncies detectades en el sèrum sanguini).
- **Informació per imatge**: Dades proporcionades a partir de l'estudi radiològic del pacient, com podrien ser el nombre de lesions o el diàmetre màxim d'aquestes.

La taula conté molts valors buits, columnes irrellevants o molt correlacionades i pacients amb poques dades, per tant, haurem de realitzar un preprocessament exhaustiu de les dades.

Explicam a continuació les variables de les que en parlarem durant la memòria:

- Informació clínica:
  - *Obesity*: Variable categòrica que representa l'abscència o presència d'obesitat en el pacient.

  - *Alcohol_abuse*: Variable categòrica que representa si el pacient és o no alcohòlic.
  - *Statin_treatment*: Variable categòrica que representa si el pacient ha estat sotmès a un tractament d'estatines o no. Les estatines són fàrmacs que redueixen els nivells de colesterol en sang.

- Informació de laboratori:

  - *Total_Bilirrubine*: Variable numèrica que mesura la quantitat total de bilirrubina en sang.

- Informació per imatge:

  - *MAX_TM_DIAMETER*: Variable quantitativa que indica el diàmetre màxim de la major de les masses tumorals del pacient.
  - *TECHNICALLY_ACCEPTABLE_FOR_LIRADS* i *IV_CONTRAST*: Variables categòriques sobre la qualitat de l'exploració radiològica.
  - *IVb*, *IVa*, *VIII*: Variables categòriques que representen localitzacions per segments del tumor.

2.3.3 Base de dades per massa tumoral

Es tracta d'una extensió de l'anterior base de dades, on estudiarem les masses tumorals de manera independent. Aquesta consta de 202 masses tumorals i 99 variables. La variable predictora és *Viable*, que serà un 0 quan el tumor es redueixi per complet, i un 1 quan no sigui així.

Aquestes variables contenen les dades demogràfiques i la informació clínica i de laboratori que teniem a la taula anterior, òbviament, masses tumorals d'un mateix pacient tindran els mateixos valors. La informació radiológica l'hem substituïda per una altra que prové també per imatge, però específica per a cada massa tumoral, com la localització (per zones del fetge) del tumor, el seu tamany, etc.

Aquesta base de dades conté també valors buits i l'haurem d'estudiar prèviament a aplicar els models de predicció, però és més completa que l'anterior.

Vegem ara algunes de les variables que tindrem en compte:

- Dades demogràfiques:

  - *AGE_AT_TACE*: Variable numèrica que indica l'edat del pacient el dia del tractament.

- Informació clínica:

  - *no_active_ex*: Variable categòrica que representa si el pacient és o ha estat fumador.

- Informació de laboratori:

  - *Potasio*: Variable quantitativa sobre el nivell de potasi.
  - *Creatinine mg/dL*: Variable quantitativa que mesura el nivell de creatinina en mg/dL.
  - *Sodio*: Variable quantitativa que mesura el nivell de sodi.

- Informació per imatge:
  - *Size*: Variable numèrica que mesura l'àrea màxima del tumor en 2D.
  - *7-11_CAT*: Criteri categòric que mesura com d'apropiat és realitzar el TACE, combinant el valor del major diàmetre amb el nombre total de tumors. Aquesta en concret no serà específica per a cada tumor, sinó per pacient.
  - *Location Observation*: Variable categòrica on cada valor representa una o més seccions del fetge, que serà on està ubicat el tumor.
  - *LR M Criteria*: Variable categòrica que ens diu si són o no algun dels tipus de tumors desdiferenciats més agressius.
  - *LI-RADS*: Variable categòrica de 8 valors que classifica segons el grau de malignitat estimat per dos radiòlegs mitjançant el diagnòstic per imatge.

CAPÍTOL 3

EXPERIMENTACIÓ I RESULTATS

3.1 Wisconsin Breast Cancer Dataset

Com s'ha explicat a la secció 2.3.1, tractam primer aquesta base de dades senzilla que condueix a bons resultats, per a poder trobar un bon model predictiu i poder estudiar si la nostra forma de caracteritzar la incertesa és efectiva, sense haver d'estar pendents de problemes complexos provocats per la base de dades.

3.1.1 Model

El nostre primer objectiu amb aquest conjunt de dades és aplicar-hi un model predictiu que ens retorni no només un nombre que ens digui com de probable és que el tumor sigui maligne, i.e *diagnosis = 1*, sinó que ens retorni un conjunt creïble que poguem analitzar. Com estam classificant una variable binària, aquest conjunt creïble no serà res més que un interval.

Per a aconseguir-ho, repetirem el nostre model moltes vegades, guardarem les probabilitats en una llista, i generarem un interval creïble a partir del conjunt de probabilitats obtingudes.

Volem extreure d'aquesta manera un interval creïble per a cada pacient, per a poder estudiar la incertesa de manera general i específica, de tal forma que amb els futurs pacients de la nostra base de dades, tinguem una representació local de com de segurs estam sobre aquell determinat pacient, ja que el mateix model pot estar molt segur sobre un pacient, i tenir una quantitat enorme d'incertesa en un altre. Tenir constància d'això en un anàlisi mèdic, on els pacients es tracten de forma individual i una mala decisió pot tenir conseqüències greus, és de vital importància. Per això, realitzarem LOOCV, escollint un pacient com a conjunt *test*, realitzant la predicció amb tots els

pacients restants com a conjunt d'entrenament, i iterant el procés per a cada un d'ells.

Amb l'objectiu de repetir el model i obtenir un gran nombre de probabilitats, realitzarem, per a cada pacient, un *bootstrap* de 1000 repeticions. Per a cada una d'elles, es fa un mostreig amb reemplaçament dels pacients restants i s'hi aplica el model. Escollim aquest nombre de repeticions per a que no sigui computacionalment massa costós, però que a la vegada el biaix del resultat no sigui significatiu.

El model escollit, al tractar amb una classificació binària, és la regressió logística. Ens queda doncs un model com el següent, on hi ficam el *data set* i ens retorna una taula on cada pacient té una llista de 1000 probabilitats:

Algorithm 1: LOOCV amb Bootstrap i Regressió Logística

Data: Taula de dades
Result: Taula de probabilitats bootstrap
Seleccionar la columna a predir: $y \leftarrow$ valors de la columna "diagnosis";
Escalar X amb un escalador de característiques $scaler$;
Inicialitzar Leave-One-Out Cross Validation (LOOCV);
Inicialitzar el model de regressió logística;
Inicialitzar llista buida per emmagatzemar les probabilitats;
for ***each*** *tumor i en LOOCV* **do**
    Dividir X_{scaled} en X_{train} i X_{test};
    Dividir y en y_{train} i y_{test};
    Inicialitzar una llista buida per guardar les probabilitats del cas concret;
    for *1000 iteracions* **do**
        Generar una mostra bootstrap de X_{train} i y_{train};
        Entrenar el model amb la mostra bootstrap;
        Obtenir la probabilitat de "diagnosis = 1"per X_{test};
        Afegir la probabilitat a la llista;
    end
    Afegir les probabilitats a la llista buida principal;
end
Convertir la llista completa en una taula de dades;

3.1.2 Avaluació model

Ens interesa ara, abans d'entrar a analitzar els intervals creïbles, veure si el model prediu bé. Per a això, calcularem la mitjana de les probabilitats de cada pacient i determinarem un llindar per a classificar amb 0 si la mitjana és inferior, i 1 si és superior a aquest.

Feim proves amb distints valors de llindars, i a simple vista sembla que obtenim millors resultats amb el 0.5. Amb aquest llindar, miram d'avaluar amb una mica més de detall com de bo és aquest model i obtenim uns grans resultats:

Taula 3.1: Matriu de confusió model 1.

	Predicció: Negatiu	Predicció: Positiu
Real: Negatiu	354	3
Real: Positiu	7	205

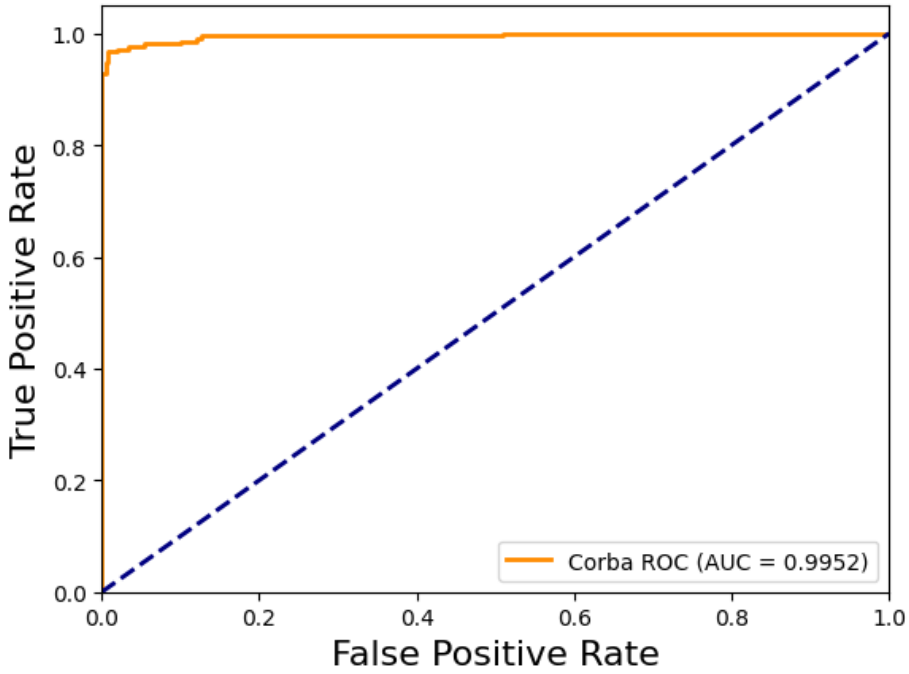
Matriu de confusió excel·lent, amb només 3 falsos positius i 7 falsos negatius, dels 569 pacients. Això condueix a uns valors altíssims en les següents mètriques de precisió:

Taula 3.2: Mètriques de precisió.

Precisió	Sensibilitat	F1-Score	Exactitud
0.99	0.97	0.98	0.98

Feim la curva ROC, que és molt bona i calculam l'AUC. Obtenim un $AUC = 0.9952$, el que ens diu que estam predint amb moltíssima precisió. Calculam l'umbral òptim i obtenim que és un 0.51, pel que suposam que amb un gruix de dades més alt, aquest valor tendirà al 0.50. Per tant, un cop obtenim l'interval creïble, predirem amb 1 els intervals on la seva mitjana sigui superior a 0.50 i amb 0 el cas contrari, i mantindrem aquest llindar durant tot el procés.

Figura 3.1: Corba ROC del model 1.



3.1.3 Anàlisi de la incertesa obtinguda

Ara ja sí, ens centram en obtenir i analitzar els nostres conjunts creïbles amb la nostra mètrica. Calculam l'interval creïble amb un grau de confiança del 95% per a cada pacient i ho guardam a una taula, per a poder calcular, per a cada pacient, la seva incertesa total (TP), l'aleatòria (AP) i l'epistèmica (EP).

Feim una anàlisi general d'estadística descriptiva sobre els resultats obtinguts, i filtram els pacients per obtenir aquells que tinguin molta o poca incertesa. D'aquesta anàlisi, extreim aquesta taula i les següents conclusions:

Taula 3.3: Incertesa model 1.

	Incertesa aleatòria (AP)	Incertesa epistèmica (EP)	Incertesa total (TP)
Mitjana	0.0103	0.0709	0.0813
Desviació típica	0.0347	0.1625	0.1860

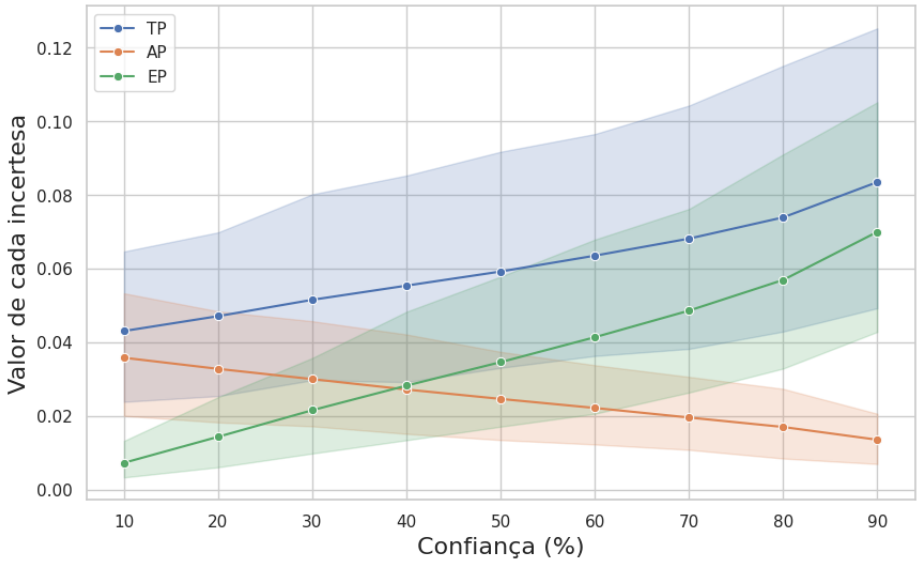
- Poca incertesa, ens equivocam poc i, a més, estam molt segurs, en general, de les prediccions que feim.
- Gran part de la nostra incertesa és epistèmica i per tant reduïble.
- Hi ha pacients on estam segurs del seu diagnostic, un pacient té una $TP = 0.0$, i d'altres dels que no en sabem res, el pacient que en té més té $TP = 0.9943$. Per tant, a l'hora d'extrapolar-ho a un cas real, hauríem de ser conscients de que no per tots els pacients obtindriem resultats útils, encara que el model en general funcioni molt bé.
- Tenim 102 pacients amb $TP > 0.1$, llavors amb tota la resta tenim un diagnostic amb molt poca incertesa i hi podriem fer prediccions fiables.
- La correlació entre TP i EP és molt alta (0.9892), llavors la incertesa total té molt que veure amb l'epistèmica, que és reduïble.

3.1.4 Intervals creïbles

Hem escollit de manera subjectiva la confiança desitjada en els nostres intervals creïbles, per tant volem ara mesurar com ens canviaria l'estudi si variam aquest grau de confiança.

Al disminuir el grau de confiança, el que estam fent és escollir una menor quantitat d'informació, deixant fora totes les probabilitats que no entren dins l'interval. De manera molt natural, veiem que com més gran és el grau de confiança, major incertesa epistèmica obtenim, ja que per definició, la longitut de l'interval serà més gran. Malgrat això, a l'estar tractant amb més dades, pujar el grau de confiança també implica una menor incertesa aleatòria. Com la incertesa epistèmica és reduïble, escollir un alt grau

Figura 3.2: Variació de la incertesa segons la confiança dels intervals.



de confiança farà que l'estudi no perdi dades pel camí i tingui el mínim d'incertesa aleatòria.

3.1.5 Renou a les dades

Vegem ara com afecta a la incertesa el renou a les dades. A la nostra base de dades original li afegim renou de manera proporcional a la desviació típica de cada columna:

$$\text{renou} \sim N(0, 2\sigma_k)$$

Entrenam el model amb un *training* on cada variable tendrà renou i un *test* no alterat. Un cop fet, replicam exactament la mateixa anàlisi per tal de cercar la diferència, que queda molt ben representada per les mitjanes de cada un dels tipus d'incertesa i el seu respectiu percentatge d'augment:

Taula 3.4: Incertesa amb renou o sense.

	Sense renou	Amb renou	Percentatge d'augment
Incertesa aleatòria (AP)	0.0103	0.0172	67%
Incertesa epistèmica (EP)	0.0709	0.0795	12,13%
Incertesa total (TP)	0.0813	0.0967	18,95%

Es pot veure com el percentatge d'augment de la incertesa aleatòria és considerablement major que el de l'epistèmica. Com a conclusió, hem vist de manera empírica que la incertesa aleatòria realment mesura la incertesa intrínseca de les dades, ja que amb unes dades alterades, per molt que milloràssim el model, tindriem una incertesa aleatòria molt major a la que tindriem amb les mateixes dades sense alterar.

3.2 Dades de l'HUSE per pacient

Amb el model a punt i la mètrica contrastada, estam preparats ja per afrontar un repte més complicat, intentar predir l'efectivitat d'un tractament a pacients reals que pateixen hepatocarcinoma. Ens referim a la base de dades de la secció 2.3.2. La classificació seguirà sent binària, i la variable predictora serà RECIST, on un 1 significarà que el pacient ha respost bé al tractament (resposta parcial o completa), i un 0 signifacarà que no (es queda igual o pitjor).

Per altra banda, ens plantejam dos nous objectius, el primer relacionat amb el preprocessament de dades. No només ens ajudaran en la predicció, sino que ens donaran informació rellevant nova sobre les dades:

- Imputació de valors buits: La base de dades presenta molts de valors buits, ja sigui per pèrdua d'informació, per manca de dades de determinats pacients, etc. El nostre primer objectiu serà determinar quan val la pena imputar aquests valors buits o, per altra banda, eliminar la variable o el pacient perquè no tenim prou informació. Finalment, en els casos on imputem, com fer-ho. Per a això, hem provat tres sistemes d'imputació diferents, de menys a més complexitat: imputació per mitjana i moda, imputació per *K-Nearest-Neigtbours* i imputació mitjançant la distància de Gower.

- Importància de les variables: A banda de predir bé, també ens interesa saber quines són les variables que ens estan donant més informació, per tal de millorar estudis posteriors, indicar quins apartats de la recollida de dades són els que se'ls hi ha de donar més importància, etc.

3.2.1 Imputació de dades

Per diversos motius, en el món de la medicina és molt habitual trobar valors buits en les bases de dades dels pacients. Per això, com tractar les variables afectades o com omplir aquests buits sintèticament, és un aspecte important.

Després d'un preprocessament damunt de la base de dades original, on eliminam columnes buides, variables molt correlacionades i pacients amb molt poques dades o sense variable predictora, seguim tenint un gruix important de valors buits. A més, trobam un segon obstacle, que és la presència de variables numèriques i la de variables categòriques, que haurem de tractar de manera diferent. Trobam tres solucions al problema, i escollirem finalment la tercera, ja que, a banda de ser la millor des d'un punt de vista teòric, és la que presenta menys incertesa a l'hora d'aplicar-hi el model.

A qualsevol dels tres experiments, renunciam a totes les variables que presentin un 35% o més de valors buits, ja que hem considerat que a partir d'aquest llindar, qualsevol tipus de manipulació sintètica pot estar greument esbiaixada.

Imputació de mitjana i moda

La primera solució és la més simple i la que computacionalment presenta menys problemes. Es tracta simplement d'estudiar cada variable de manera independent, i omplir els buits de la següent manera:

- Per a les variables quantitatives, imputar la mitjana de tots els valors reals a cada buit.
- Per a les variables categòriques, imputar la moda dels valors reals a cada buit.

Aquesta solució no altera la mitjana, que seguirà sent la mateixa, però sí la seva desviació típica, que es reduirà bastant. A més, serà molt sensible als valors atípics. En el cas de les categòriques, presentarà biaix, especialment si no hi ha molta diferència entre el nombre de vegades que surt un valor o l'altre.

Malgrat això, al ser un procés tan simple i interpretable, no hagués estat una mala opció en el cas d'haver tingut pocs valors buits o uns resultats similars al final del procés.

Imputació per K-nearest neighbors

Una segona solució serà la d'aplicar el mètode d'imputació per K-Nearest Neighbors. És un mètode molt extés, útil i fàcil d'usar. Malgrat això, no es pot aplicar a dades categòriques, per tant, haurem d'aplicar el mateix model d'imputació per moda en les categòriques, i tractar les numèriques apart. El procés per a les numèriques és també senzill:

1. Es calcula la distància euclídea entre parells de pacients, obviant les variables on un dels dos té algun valor buit.
2. Quan trobam un valor buit, es seleccionen els K (en el nostre cas, 5) veïnats més propers al pacient en concret.
3. S'imputa la mitjana dels valors per aquesta variable dels 5 pacients seleccionats.
4. S'itera el procés fins a tenir la taula de dades completa.

D'aquesta manera, el procés ja no és tan sensible als valors atípics, i s'imputa un valor sintètic que en principi ha d'estar més proper al valor real. De fet, hi trobam millores en la incertesa del model.

Malgrat aquestes, els veïnats propers es troben només amb l'informació que ens donen les variables numèriques, sense tenir en compte tota la que ens proporcionen les variables categòriques, que en són moltes en el nostre cas.

Imputació usant la distància de Gower

Finalment, hem trobat una solució molt semblant a l'anterior, però que ens ha permés tractar també les variables categòriques i tenir-les en compte a l'hora de trobar els veïnats propers. La distància entre els pacients es calcula mitjançant la distància de

Gower, explicada en la secció 2.1.6, on hi intervenen ambdós tipus de variables.

No obstant, aquesta distància té un problema, i és que no es pot calcular amb valors buits. Per tant, hem hagut d'afegir primer dades sintètiques als valors buits, mitjançant el mètode de la mitjana i la moda, per a poder calcular les distàncies i després poder corregir aquests valors. Això, evidentment, ens genera un biaix, però com que obtindrem menys incertesa al final del procés comparat amb el segon mètode, ens decantarem per aquest camí.

1. Imputam dades sintètiques als valors buits mitjançant el primer mètode, senzill computacionalment.
2. Amb les dades sense valors buits, calculam una matriu de distàncies amb la distància de Gower, guardant així les distàncies entre tots els pacients.
3. Treballam ara amb la base de dades amb valors buits. Per a cada valor buit trobat, es seleccionen els 5 pacients més propers en la matriu anterior, excloent-se a sí mateix, i:
    - Imputam la mitjana dels 5 valors si la variable és numèrica.
    - Imputam la moda dels 5 valors si la variable és categòrica.

Aquesta forma d'imputar les dades no és sensible als valors atípics, té en compte tota la informació a l'hora de torbar pacients semblants i ens permet imputar dades sintètiques tant numèriques com categòriques. A més, aplicant el model anterior, és on hi trobam millors resultats.

3.2.2 Importància de les variables

És important per a l'estudi clínic general, no només pel nostre treball de recerca, saber quines variables tenen pes en la predicció, per a poder donar informació als metges i conduir en aquella direcció les recerques posteriors. És per això que hem cercat una manera de quantificar, dins del nostre model, aquesta importància.

Al tenir tantes variables, ens ha resultat adequat aplicar una penalització LASSO dins la nostra regressió logística. Aquesta penalització fa que el coeficient de regressió de les variables amb poc pes sigui zero, i d'aquest manera, directament no les tinguem en compte a l'hora de predir.

Per quantificar la importància de cada variable, hem inicialitzat un contador abans d'aplicar el model, que dóna el percentatge de vegades que aquesta variable ha estat considerada important, és a dir, de totes les iteracions que fa el model, en quantes d'elles el seu coeficient no ha estat zero.

D'aquesta manera, podrem dir que les variables amb un percentatge pròxim al 100% són rellevants per a l'estudi, mentre que aquelles que surten poques vegades, no ens aporten gaire informació.

3.2.3 Model amb regressió logística

Ens enfocam ara ja sí en fer prediccions amb la base de dades ja filtrada i sense valors buits, completada mitjançant la distància de Gower, i tenint en compte que voldrem, a més, tenir una referència quantitativa de com d'important és cada variable. Per això, realitzarem el mateix model de regressió logística, aquest cop amb penalització LASSO, i hi aplicarem el *bootstrap* i el *LOOCV* per a poder obtenir els intervals creïbles.

Els resultats que obtenim no són gaire bons, amb una incertesa mitjana de 0.6456, pràcticament tota epistèmica, ja que només en tenim 0.0031 d'aleatòria. A l'observar la matriu de confusió de la nostra classificació, és quan ens adonam d'un dels principals problemes d'aquesta base de dades és que està molt desequilibrada, amb 110 pacients on la seva classificació real és positiva, el tractament funciona, i només 21 casos on no. El model, influenciat per això, no classifica de manera correcta cap negatiu, oferint-nos aquesta matriu de confusió tan desequilibrada:

Taula 3.5: Matriu de confusió model amb regressió logística.

	Predicció: Negatiu	**Predicció: Positiu**
Actual: Negatiu	0	21
Actual: Positiu	9	101

Davant aquests resultats, ens veiem obligats a corregir aquest desequilibri. És per això que empleam la tècnica de SMOTE, explicada en la secció 2.1.5, per tal de generar sintèticament mostres al *training* de la classe amb menys representants, entrenar el model, i posteriorment fer la classificació.

Abans de provar-ho, i encara que no poguem extreure massa conclusions d'aquest experiment, provarem el model amb SMOTE al training, per comprovar que tingui una certa coherència. Els resultats són molt bons, amb una precisió pròxima al 100%, i es per això que aplicam el nostre model amb SMOTE:

Algorithm 2: LOOCV amb SMOTE, Bootstrap i Regressió Logística

Data: Dataset X, y, model de regressió logística, nombre d'iteracions
Result: Llista de probabilitats i comptador de variables seleccionades
for ***each*** *pacient i en LOOCV* **do**
  Dividir X_{scaled} en X_{train} i X_{test};
  Dividir y en y_{train} i y_{test};
  Aplicar SMOTE per balançar la classe:
  $(X_{train\_smote}, y_{train\_smote}) \leftarrow SMOTE(X_{train}, y_{train})$;
  Inicialitzar una llista buida;
  for ***each*** *iteració de bootstrap fins al nombre total d'iteracions* **do**
    Generar una mostra bootstrap de X_{train_smote} i y_{train_smote};
    Entrenar el model amb la mostra bootstrap;
    Obtenir la probabilitat de "diagnosis = 1"per X_{test};
    Afegir la probabilitat a la llista;
    Actualitzar el comptador de variables seleccionades;
    $variable\_counts \leftarrow variable\_counts + (coeficients\_model \neq 0)$;
  end
  Afegir les probabilitats a la llista
end

Els resultats milloren els del model anterior sense SMOTE. La nova matriu de confusió demostra que s'ha corregit el problema del desequilibri i que ara ens equivocam menys:

Taula 3.6: Matriu de confusió model amb SMOTE.

	Predicció: Negatiu	**Predicció: Positiu**
Actual: Negatiu	6	15
Actual: Positiu	36	74

Malgrat això, com abans el model classificava positivament a quasi tots els pacients de manera molt segura, i ara hem corregit el desequilibri, la incertesa ha augmentat. Així i tot, la aleatòria disminueixa menys de la meitat, fent decréixer la incertesa irreduïble del model.

Taula 3.7: Incertesa amb SMOTE i sense.

	Sense SMOTE	Amb SMOTE	Percentatge d'augment
Incertesa aleatòria (AP)	0.0026	0.0012	-53.85%
Incertesa epistèmica (EP)	0.6624	0.7304	10.27%
Incertesa total (TP)	0.6650	0.7316	10.02%

Les 10 variables que tenim en compte en un nombre major de prediccions són aquestes, explicades totes a la secció 2.3.2:

Taula 3.8: Variables importants de la taula de dades per pacient.

Variable	**Freqüència (%)**	**Tipus**
Obesity	96.506107	Clínic
MAX_TM_DIAMETER	96.175573	Radiol
TECHNICALLY_ACCEPTABLE_FOR_LIRADS	94.820611	Radiol
Total_Bilirrubine	91.747328	Clínic
Statin_treatment	91.543511	Clínic
IVb	90.174809	Radiol
IV_CONTRAST	89.709160	Radiol
IVa	89.112214	Radiol
Alcohol_abuse	87.964885	Clínic
VIII	86.771756	Radiol

3.2.4 Model amb random forest

Davant una no molt bona predicció amb el model de regressió logística, provam un nou enfoc per veure si hi trobam millores. Realitzarem ara el mateix experiment, on l'únic canvi és el model de classificació.

Com hem fet anteriorment, ho provam primer en el *training*, i com obtenim també una precisió pròxima al 100%, duim a terme el nostre model.

Hem realitzat aquest cop un *grid search*, explicat en la secció 2.1.4, per a trobar els hiperparàmetres adequats i realitzar el millor *random forest* possible.

Taula 3.9: Comparació incertesa entre els dos models.

	RL amb SMOTE	Random Forest	Percentatge d'augment
Incertesa aleatòria (AP)	0.0012	0.2074	+17283.33%
Incertesa epistèmica (EP)	0.7304	0.1618	-77.86%
Incertesa total (TP)	0.7316	0.3691	-49.56%

Els resultats obtinguts milloren les mètriques d'incertesa total i epistèmica, i empitjoren molt (el primer valor era molt baix) els de la aleatòria. Malgrat aquesta teòrica millora en la incertesa total, és quan veiem la matriu de confusió que ens adonam que la tècnica de SMOTE i l'hiperparàmetre *classweight* = *balanced*, no han estat suficientment fortes com per compensar el desequilibri en les dades i el que està fent el model és classificar-ho quasi tot com a positiu, com el primer model de regressió logística sense SMOTE:

Taula 3.10: Matriu de confusió model amb random forest.

	Predicció: Negatiu	**Predicció: Positiu**
Actual: Negatiu	0	21
Actual: Positiu	3	107

Per tant, definim el model de regressió logística amb penalització LASSO i SMOTE com el millor per a aquesta base de dades.

3.3 Dades de l'HUSE per massa tumoral

La investigació, per diversos motius, ens condueix a treballar les dades amb un enfoc diferent. A partir d'ara, no treballarem per pacient, sinó per massa tumoral, és a dir, estudiarem cada massa tumoral de manera independent, ja que tenim pacients amb diverses masses tumorals on cada una d'elles ha respost al tractament de manera diferent. La variable predictora ara serà *Viable*, que és un 0 quan el tumor s'ha reduït per complet, i un 1 quan no ha estat així.

A més, una petita part dels valors buits ha estat completada desde l'hospital, un fet que ens ajuda a haver d'afegir menys dades sintètiques i per tant reduir la incertesa aleatòria.

3.3.1 Preprocessament

Una vegada carregada la base de dades actualitzada, la part del preprocessament més costosa ha estat la de generar una fila per a cada massa tumoral, copiar la informació demogràfica, clínica i de laboratori que compartiran totes les que siguin d'un mateix pacient, i afegir-hi la informació tumoral per imatge corresponent.

Després hem seguit el mateix procediment que en l'anterior experimentació:

1. Eliminar les columnes que tenguin més d'un 35% de valors buits.

2. Imputar els valors buits restants mitjançant el mètode que usa la distància de Gower, descrit anteriorment.

Un aspecte important a considerar d'aquesta nova base de dades és que ja no tenim el desequilibri de classes anterior. A l'estudiar les masses tumorals una a una, tenim 107 casos positius i 96 de negatius, i aquesta diferència no ens suposarà un problema, renunciant així a la tècnica de SMOTE.

3.3.2 Random Forest

Malgrat el poc èxit obtingut prèviament amb aquest model, el tornarem a provar amb la nova base de dades, ja que el que ens donava complicacions era, precisament, aquest desequilibri de classes que ja no tenim.

Calculam els hiperpàmetres òptims del *random forest* mitjançant un *grid search* i entrenam el nostre model com sempre, amb el *bootstrap* i el LOOCV per a extreure moltes probabilitats, trobar un interval de confiança i estudiar la incertesa.

Els resultats són molt millors que el mateix model amb les altres dades, oferint una matriu de confusió com la següent i predint correctament el 68% de les masses tumorals:

Taula 3.11: Matriu de confusió model amb random forest.

	Predicció: Negatiu	**Predicció: Positiu**
Real: Negatiu	57	39
Real: Positiu	26	81

Les mitjanes de la incertesa són:

Taula 3.12: Incertesa del model amb random forest.

	Incertesa aleatòria (AP)	Incertesa epistèmica (EP)	Incertesa total (TP)
Mitjana	0.1937	0.3842	0.5780

En aquest cas, predim pràcticament igual de bé que amb la regressió logística, però com que obtindrem una mica menys d'incertesa, analitzarem els resultats més a fons en el segon cas.

3.3.3 Regressió logística

Aplicam el model de regressió logística amb *bootstrap* i LOOCV, aquest cop sense la tècnica de SMOTE. Al tenir un conjunt de dades actualitzat amb menys valors buits i no tenir desequilibri de classes, veurem que els resultats són molt millors.

Cal destacar que, a l'hora de fer el *grid search*, obtenim que l'hiperparàmetre òptim és $\lambda = 10$, on λ és el paràmetre de penalització LASSO, respecte al $\lambda = 1$ que obteniem a la base de dades per pacients. Això el que ens indica és que serem molt més restrictius a l'hora d'incloure o no una variable, el que es traduirà amb major incertesa aleatòria (ja que tenim en compte menys dades) però també amb una selecció de variables molt més explicativa i una millora en les prediccions.

La matriu de confusió és molt semblant a la del *random forest*, on només ens equivocam predint una vegada més:

Taula 3.13: Matriu de confusió del model amb random forest.

	Predicció: Negatiu	**Predicció: Positiu**
Real: Negatiu	52	44
Real: Positiu	22	85

Els resultats en relació a la incertesa són també semblants, amb un increment important de la incertesa aleatòria justificat per la restricció LASSO:

Taula 3.14: Incertesa del model amb random forest.

	Incertesa aleatòria (AP)	Incertesa epistèmica (EP)	Incertesa total (TP)
Mitjana	0.2355	0.3289	0.5644

Les variables que han donat més informació en les prediccions són les següents, explicades a 2.3.3:

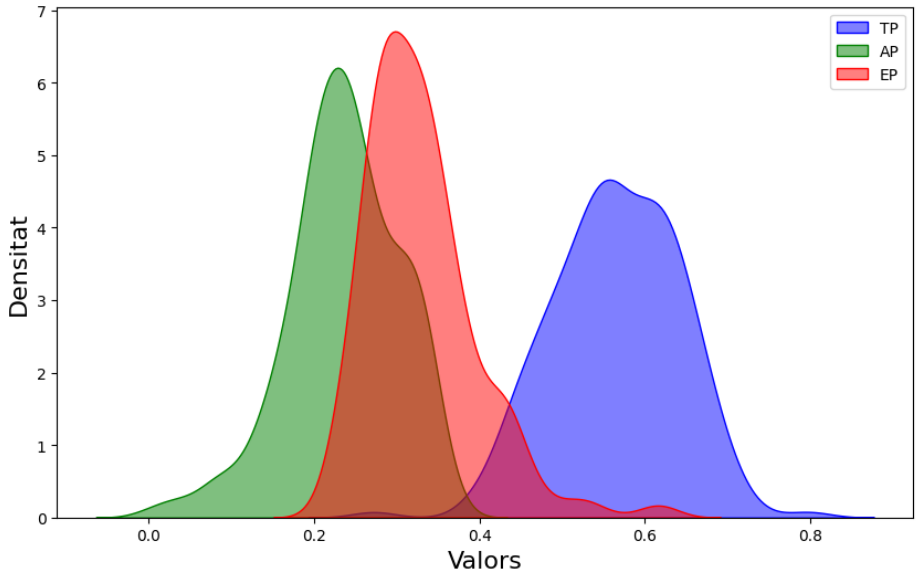
Taula 3.15: Variables importants de la taula de dades per massa tumoral.

Variable	**Freqüència (%)**	**Tipus**
Size 2D mm.1	94.168159	Radiol
7-11_CAT	89.703483	Radiol
Location Observation	83.613433	Radiol
Potasio	82.261194	Clínic
LR M criteria	75.329353	Radiol
AGE_AT_TACE	73.469652	Clínic
Creatinine mg/dL	52.886567	Clínic
no_active_ex	49.964179	Clínic
Sodio	42.806965	Clínic
LI-RADS	36.668657	Radiol

3.3.4 Anàlisi de la incertesa

Els tres valors de la incertesa ens donen poca informació en aquest cas, ja que, com es veu representat al gràfic de densitat, cada una d'elles està molt concentrada al voltat del seu valor mitjà, per tant, no podrem dir com abans que per algunes masses tumorals la predicció és segura mentres que d'altres no en sabem res. En el nostre cas, cap predicció és molt segura però a totes les prediccions tenim un cert grau de confiança.

Figura 3.3: Dsitribucions de AP, EP i TP.



Tant sols hem trobat una sola massa tumoral que tingui una incertessa total per davall del 0.3, mentre que tant sols una d'elles té una incertesa total per damunt de 0.75. La resta es mantenen dins aquest marge, que no ens assegura que cap predicció sigui ni bona ni dolenta.

A més, la incertesa en les prediccions correctes és una mica menor que els prediccions errònies, encara que la diferència no és gens significativa.

CAPÍTOL 4

CONCLUSIONS

L'objectiu principal del treball era caracteritzar la incertesa que presentava el nostre estudi. En un context complicat, on fer bones prediccions ha estat condicionat per l'alta complexitat del problema i la manca d'informació, hem volgut definir el millor possible la quantitat d'incertesa obtinguda i el seu tipus, per tenir constància de si aquesta és o no reduïble.

A banda d'això, també hem partit amb l'objectiu de contribuir al projecte de l'HUSE amb tres aspectes molt concrets: un mètode per a imputar els valors buits afegint la menor incertesa possible, un mètode per extreure informació sobre quines variables són més importants, i un model predictiu de referència.

Malgrat haver millorat les prediccions fetes a altres projectes anteriors, com els treballs de recerca de na Maria Antònia Colomar [9] i de na Maria del Mar Deyà [10], no hem trobat un model que presenti poca incertesa. Però això no ens ha impedit poder fer una proposta de solució als tres problemes esmentats anteriorment.

Primer, analitzarem el nostre mètode de quantificació i caracterització de la incertesa, ja que hi hem trobat tant aspectes positius com negatius. A més, explicarem en detall la informació extreta sobre les variables significatives.

Com es pot veure a la memòria, el treball de recerca en general, i el compliment dels objectius en concret, ha anat desenvolupant-se de manera progressiva, solucionant els problemes que han anat sorgint un a un. Per això, hem trobat necessària fer una recopil·lació de totes les contribucions realitzades.

Finalment, resumirem les limitacions del nostre treball i les possibles línies de treball futures.

4.1 Avaluació del mètode de caracterització de la incertesa

El nostre mètode presenta una manera clara i molt simple de representar la incertesa, i dividir-la en aleatòria i epistèmica. Malgrat això, durant l'experimentació, resumida a la taula 1.1, hem anat trobant algunes limitacions:

- **Escala desigual de les mètriques:** És un problema de definició, que ja remarquen i justifiquen els seus autors [7]. La incertesa aleatòria només es pot trobar en els valors compresos entre 0 i $\frac{1}{2}$, mentre que tant l'epistèmica com la total, entre 0 i 1. Això fa que els canvis que pugui sofrir l'epistèmica es vegin molt més representats en la incertesa total, que els que pugui tenir l'aleatòria. Es per aquest motiu que hem intentat recollir sempre el percentatge d'augment durant el transcurs del treball.

- **Separació imperfecta:** Malgrat les fórmules intentin distingir completament la incertesa, en problemes complexos com el nostre, aquesta separació no és del tot precisa, ja que ambdós tipus d'incertesa estaran correlacionats.

- **Desequilibri de classes:** Males prediccions poden conduir molts cops a incerteses baixes per diverses raons. Dins el context del nostre treball, hem pogut veure a la secció 3.2.4 que un model esbiaixat completament pel desequilibri de les dues classes és un mal model, però al tenir pocs errors presenta també incerteses molt baixes.

- **Incertesa en errors i encerts similars:** Ens hagués agradat poder concloure que la quantificació ens pot donar una idea de si el model s'equivocarà o no per a un pacient concret, però hem vist a la secció 3.3.4 que les vegades on el model encerta, encara que la incertesa sigui un poc més baixa, no dista molt de la incertesa obtinguda quan el model classifica malament.

Així i tot, diversos experiments ens condueixen a pensar que sí que pot ser un mètode de caracterització efectiu en molts aspectes:

- **Intervals creïbles:** En la secció 3.1.4 podem veure clarament com, al escollir treballar amb menys informació, creix naturalment la incertesa aleatòria de manera significativa.

- **Renou a les dades:** De la mateixa manera, a la secció 3.1.5, el fet d'afegir renou a les dades incrementa la incertesa. No només això, sinó que l'aleatòria creix molt més que l'epistèmica.

- **LASSO:** En la secció 3.3.3, augmentam el valor de λ en la restricció LASSO, que fa que el model sigui molt més restrictiu i molts més coeficients siguin 0, provocant així que un menor nombre de variables intervenguin en la predicció. Podem veure com, malgrat sigui un model amb bastanta menys incertesa que els anteriors, els valors de la incertesa aleatòria es disparen de manera natural.

- **SMOTE:** En la secció 3.2.3, podem veure que el fet d'aplicar la tècnica de SMOTE, que ens corregeix el desequilibri de classes, fa que la incertesa aleatòria disminueixi, encara que la epistèmica creixi.

4.2 Variables importants

Per respondre a un dels objectius del treball, hem estat en comunicació directe amb la Dra. Gemma Sempere, radiòloga de l'HUSE i responsable de l'estudi clínic, per veure quines conclusions podem extreure de les variables importants obtingudes.

Estudi per pacients

A la taula 3.8 hi trobam les 10 variables que intervenen més cops en les nostres prediccions. Podem veure que tenim 4 variables de tipus clínic i 6 de tipus radiològic.

- Les 4 variables clíniques ja eren considerades rellevants en quant a la recerca feta fins el moment, per tant, aquests resultats serveixen per a reforçar aquesta idea i poder donar-los com a bons indicadors.
- De les altres 6, n'haurem de fer casos:
  - Les variables *TECHNICALLY_ACCEPTABLE_FOR_LIRADS* i *IV_CONTRAST* són variables que depenen de la qualitat de l'exploració radiològica i no del pacient en concret, per tant, no té sentit que apareguin com a bons predictors.
  - El diàmetre tumoral *MAX_TM_DIAMETER* és un bon predictor. Ja ho estava considerat per altres estudis, i el nostre ho reforça.
  - Les variables *IVb, IVa* i *VIII* són localitzacions per segments del tumor. Fins ara no es tenia constància de que hi pugui haver una relació amb l'efectivitat del tractament. Hi podria existir una explicació en base a l'anatomia vascular tumoral depenent de la seva localització, que justificàs diferències en l'efectivitiat, per duplicitat d'aport arterial o per dificultat d'accés al tractament. L'aparició d'aquestes variables al nostre estudi pot obrir una línia d'investigació nova.

Estudi per massa tumoral

A la taula 3.15 podem trobar les 10 variables que intervenen més cops en el segon estudi, que té millors resultats que l'altre. Hi trobam 5 variables de tipus clínic i 5 de tipus radiològic.

- Les 5 de tipus clínic ja estaven considerades com a bons predictors, per tant contrastam el que ja sabíem.
- De les radiològiques tornam a fer casos:
  - La variable *Size*, és la que representa la mida de la massa tumoral, i és evident que serà bona predictora, correspon a la variable *MAX_TM_DIAMETER* a la taula 3.8, per tant, podem concloure que serà realment una variable important a l'hora de fer prediccions.
  - *7-11_CAT*, *LR-M* i *LIRADS* són graus de càrrega tumoral i de tumors diferenciats més agressius, i ja teníem constància de que eren rellevants.

– *Location Observation* segueix amb la mateixa línia del que hem comentat abans, representa la localització de la massa tumoral, que en principi no es tenia constància de la seva importància, i per segon cop obtenim que es rellevant, també en aquest estudi.

Variables no seleccionades

Algunes variables que es consideraven importants a la literatura no han sortit al nostre estudi, posant així en dubte la seva importància real. Alguns exemples són:

- Els *Scores* sobre la gravetat del tumor.
- Algunes variables de laboratori com: *plaquetes*, *GGT*, *AST*, *Quick(%)*, *Creatinine* o *Albumine*.

4.3 Contribucions

A mode de resum, ennumeram totes les contribucions realitzades durant el transcurs del treball:

1. **Caracterització de la incertesa:** Aplicam i analitzam el mètode de quantificació de la incertesa proposat per Hullemeier [7] a partir dels intervals creïbles. Som capaços de quantificar-la i separar-la en aleatòria i epistèmica.

2. **Imputació de valors buits:** Mètode d'imputació de valors buits basat en la distància de Gower i els K-nearest-neighbours. Concloem que aquest mètode és un poc menys incert que altres mètodes tradicionals menys complexes.

3. **Obtenció d'intervals creïbles:** Mitjançant LOOCV i el bootstrap, trobam la forma d'obtenir moltes probabilitats del mateix succés amb les que poder obtenir un interval creïble i poder estudiar la seva incertesa a posteriori.

4. **SMOTE:** Proposam una solució al problema del desequilibri de classes.

5. **Variables importants:** Mitjançant la penalització LASSO, trobam la manera de conèixer quines variables donen informació al model més cops, amb l'objectiu de poder classificar les variables més importants. A partir d'això, hem pogut donar informació pràctica als responsables de l'estudi clínic.

6. **Model de regressió logística i de random forest:** Encara que els resultats predictius dels nostres dos models no siguin molt bons degut a la complexitat del problema, milloren els resultats obtinguts als estudis prèvis.

7. **Canvi d'enfoc a masses tumorals:** Hem pogut concloure que el canvi d'enfoc de l'estudi, passant d'estudiar els pacients a estudiar les masses tumorals de manera independent, presenta millors resultats, millorant així la capacitat predictora del models.

4.4 Limitacions del treball

Tot i les contribucions que hem realitzat, els resultats predictius generen una alta incertesa, degut principalment a dos motius:

- **La manca de dades:** Tant la gran quantitat de valors buits com la poca quantitat de mostres.
- **La complexitat del problema:** El tenir tantes variables de distints tipus, sense conèixer si són o no rellevants, juntament amb la variablitat de la variable predictora.
- **L'enfoc de la variable predictora en el darrer estudi:** La variable predictora *Viable*, que retorna un 0 quan el tumor es redueix al 100% i un 1 quan no és així, no permet que els tumors que s'han reduït però no el suficient per arribar al 100% contin com a èxits, sino com a fracasos, fet que pot confondre el model i produir incertesa.

A més, durant el transcurs del treball, hem aplicat tècniques computacionalment costoses, com el bootstrap, el LOOCV, la distància de Gower o l'SMOTE. Pel nostre estudi no ha suposat cap problema, però ho podria arribar a ser si ho volguéssim generalitzar a altres taules de dades més extenses.

4.5 Propostes de millora

Hem fet molta feina a partir del nostre mètode de caracterització de la incertesa però no n'hem provat d'altres. Seria una línia de recerca molt interesant realitzar els mateixos estudis amb un altre mètode i fer les respectives comparacions per veure quin d'ells s'adapta millor al nostre estudi.

Per tal de millorar les prediccions, es podrien trobar altres models més complexos per veure si arriben a millors resultats. A més, la informació que tenim a posteriori sobre la importància de les variables podria ser interesant que s'aprofitàs per a millorar la base de dades: trobar noves mostres, eliminar variables que no donen informació, i completar i millorar les variables que hem vist que sí són bones predictores.

BIBLIOGRAFIA

[1] T. H. R. T. J. T. Gareth James, Daniela Witten, *An Introduction to Statistical Learning with Applications in Pyhton.* Springer, 2023. 2.1

[2] L. O. W. K. Nitesh V.Chawla, Kevin W.Bowyer, "Smote: Synthetic minority oversampling technique," 2002. 2.1

[3] J. Gower, "A general coeficient of similarity and some of its properties," *Biometrics, Vol.27, No.4*, 1971. 2.1

[4] Wikipedia contributors, "Receiver operating characteristic," 2024, [Online; accessed 26-February-2025]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=1263557992 2.1

[5] C. R. Fox and G. Ülkümen, "Distinguishing two dimensions of uncertainty," 2011. 2.2

[6] E. H. Paul Hofman, Yusuf Sale, "Quantifying aleatoric and epistemic uncertainty with proper scoring rules," 2024. 2.2.3

[7] M. H. S. Eyke Hüllemeier, Sebastian Destercke, "Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison," 2022. 2.2.3, 2.2.4, 4.1, 1

[8] O. L. M. W. Nick Street, William H. Wolberg, "Nuclear feature extraction for breast tumor diagnosis," 1992. 2.3.1

[9] M. A. C. Riutort, "Característiques radiòmiques. extracció i anàlisi," 2023. 4

[10] M. del Mar Deyà Torrens, "Anàlisi de l'aportació de les característiques radiòmiques en estudis clínics," 2024. 4