Guia sobre l'explicabilitat en la Intel·ligència Artificial

Programa Salut/IA

TIC Salut Social

Generalitat de Catalunya

Aquest informe és fruit de l'Àrea d'Intel·ligència Artificial de la Fundació TIC Salut Social

Autors: Susanna Aussó, Didier Domínguez i Mariona Quintana.

Edició electrònica: desembre de 2022

/Índex de contingut

/Índex de figures

1.
Introducció

El Programa per a la promoció i desenvolupament de la Intel·ligència Artificial al Sistema de Salut de Catalunya té com a finalitat crear un entorn que faciliti el desenvolupament i la implementació de solucions d'Intel·ligència Artificial (IA) per a l'optimització de processos i recursos al sistema sanitari català en benefici de la societat, pacients i personal sanitari.

La Fundació TIC Salut Social ha creat aquesta guia amb l'objectiu de donar suport als actors involucrats en el desenvolupament de codi d'algorismes d'Intel·ligència Artificial aplicats a l'àmbit de la salut. El present document incideix en la importància de l'explicabilitat en aquests desenvolupaments i pretén enumerar i classificar les principals tècniques existents en funció del tipus de resultats a explicar.

L'avenç de la IA, que consisteix en la creació de sistemes capaços de raonar com l'ésser humà que aprenen de l'experiència, esbrinen com resoldre problemes davant d'unes condicions donades, contrasten informació i duen a terme tasques lògiques en tots els àmbits de la societat, és una realitat. Gràcies a la creixent disponibilitat de registres electrònics de salut, proves d'imatge mèdica digital, dades òmiques i la gran llista de conjunts de dades relacionats amb la salut, la IA també té un gran potencial per millorar el benestar de les persones [1].

Els mètodes d'aprenentatge automàtic, i més específicament les tècniques d'aprenentatge profund, s'utilitzen per crear complexos algorismes d'IA que puguin donar resposta a la necessitat d'aprendre a partir de la gran diversitat de fonts de dades de salut [2]. Però la utilització d'aquest tipus de tècniques té una contrapartida des del punt de vista del nivell de comprensió del funcionament intern dels algorismes creats. La complexitat de les xarxes neuronals artificials utilitzades fan que els mecanismes de decisió sovint siguin desconeguts fins i tot per part de qui desenvolupa els algorismes. Així doncs, cal plantejar-se algunes preguntes: som capaços d'entendre per què els models ens proporcionen una predicció determinada? En quines àrees s'estan fixant els algorismes durant el procés d'aprenentatge? Són prou automàtics o necessiten implicació humana? Tots aquests aspectes estan recollits en el document de Directrius ètiques per a una IA fiable de la Comissió Europea [3].

Així doncs, assegurar l'explicabilitat dels algorismes d'IA és clau per poder dur a terme la implantació de forma generalitzada d'aquest tipus d'eina en la pràctica clínica diària. La confiança dels i les professionals de la salut envers aquestes solucions d'IA que els han de donar suport és primordial i s'ha de construir sobre principis com la transparència i el rigor. La salut és un domini de reptes científics, però també ètics i legals, ja que les decisions preses tenen un impacte immediat en el benestar i la vida de tothom [1]. Per això la fiabilitat de les eines d'IA s'ha de construir sobre tres components [3]:

1. **IA legal,** complint totes les lleis i regulacions aplicables.
2. **IA ètica,** assegurant els principis i valors ètics.
3. **IA robusta,** tant des d'una perspectiva tècnica (garantint la solidesa de les solucions) com social (tenint en compte l'entorn en què operen).

Cadascun d'aquests conceptes és necessari, però no suficient, per aconseguir el que coneixem com a IA fiable (trustworthy AI).

Un altre aspecte a destacar és la importància de la implicació dels humans durant el procés de desenvolupament d'aquest tipus d'eina. Malgrat que es presentin i implementin tècniques d'IA cada cop més potents per resoldre problemes del món real, actualment no es tracta de sistemes totalment autònoms pel que fa a la presa de decisions, especialment en aplicacions mèdiques, on és imprescindible que els humans estiguin implicats en el procés. [4] Aquesta idea es recull en els models Human-in-the-loop (HITL) i Human-on-the-loop (HOTL), que aprofiten els punts forts dels humans i els de les màquines per produir els resultats més òptims. Els humans són els agents que poden diagnosticar com i per què els mètodes d'IA no són exitosos i revelar els seus inconvenients. D'aquesta manera estan involucrats en un procés de retroalimentació continu [5]:

- Els humans han de proporcionar dades de qualitat per tal que l'algorisme aprengui de la forma més adequada. En aquesta fase recaurien processos com l'etiquetatge de les dades, el control i mitigació dels biaixos, etc. L'algorisme d'aprenentatge automàtic aprendrà a prendre decisions a partir d'aquestes dades.
- Els algorismes sintetitzen el model per inferir allò que han après. Aquest pas pot succeir de maneres diferents, sovint de forma opaca per a qui desenvolupa l'algoritme. És en aquest punt que és important introduir eines d'explicabilitat perquè l'humà interpreti el mecanisme de presa de decisions de l'algorisme i analitzi els resultats.
- Les persones han de provar i validar el model qualificant els seus resultats, especialment en llocs on l'algorisme no està completament segur o està massa segur d'una decisió incorrecta.

2.

Intel·ligència Artificial explicable (XAI)

La Intel·ligència Artificial explicable (en anglès, explainable AI o XAI) permet que els resultats de d'un algorisme d'IA puguin ser entesos pels humans, en contraposició al concepte de "caixa negra", en el qual no és possible saber quins mecanismes s'han accionat per donar una resposta concreta (output) davant d'una entrada (input) [6].

2.1. Interpretabilitat vs. explicabilitat

Interpretabilitat: capacitat implícita d'un sistema que li permet ser lògic als ulls de les persones que el miren. La interpretabilitat indica en quina mesura un model d'aprenentatge automàtic pot associar una causa a un efecte. Així doncs, permet observar aquesta relació causa-efecte, però no dona informació sobre quins paràmetres hi intervenen.

Explicabilitat: capacitat activa d'un sistema d'executar accions que detallin el seu funcionament intern. L'explicabilitat té a veure amb la capacitat dels paràmetres d'un model, sovint ocults a les xarxes profundes, per justificar els resultats. El model pot ser explicat en termes humans, considerant tant el resultat com tot el procés intern de presa de decisions. Així doncs, l'explicabilitat dona informació sobre les característiques que han intervingut en una predicció.

2.2. Beneficis de l'ús d'eines d'explicabilitat en IA

Els beneficis de l'explicabilitat s'han d'analitzar des de múltiples punts de vista, ja que al llarg del cicle de vida d'un algorisme intervenen diversos actors sobre els quals s'observaran diferents implicacions. Així doncs, la XAI no és una qüestió purament tecnològica, sinó que invoca una sèrie d'aspectes mèdics, legals, ètics i socials [12].

2.2.1

Àmbit de desenvolupament d'algoritmes

Des del punt de vista del desenvolupament, l'explicabilitat és útil perquè els desenvolupadors puguin comprovar els seus models d'IA més enllà del simple rendiment, de manera que pot servir per detectar quan el rendiment de la predicció es basa en metadades en lloc de les dades en si.

La complexitat de les xarxes neuronals artificials fa que sovint els desenvolupadors d'un algorisme d'aquestes característiques desconeguin els mecanismes interns del model a l'hora de fer una predicció. En aquests casos la seva tasca se centra en la millora de les mètriques mitjançant la configuració de paràmetres i hiperparàmetres. La XAI permet millorar el coneixement sobre el funcionament intern de l'algorisme, i aquest coneixement més profund possibilita la comprovació sobre quines característiques estan intervenint en el resultat i, per tant, facilita la millora del rendiment de l'algorisme. En definitiva, l'explicabilitat permet desenvolupar algorismes més acurats.

Exemple: en una solució de predicció de pronòstic de COVID-19 a través de radiografies de tòrax, l'algorisme, per tal de classificar, es pot fixar en la diferència visual entre una radiografia convencional i una radiografia portàtil (que se sol fer a persones malaltes amb mobilitat reduïda com les que estan a l'UCI). En aquest cas, tot i que el rendiment del model pot ser bo, degut a la correlació entre el tipus de radiografia i l'estat del o la pacient, seria convenient revisar els mecanismes d'aprenentatge.

Aquest tipus de fenomen és conegut com a Efecte Clever Hans. Descrit inicialment en estudis científics socials, es produeix quan un experimentador o experimentadora actua sense proposar-s'ho sobre l'individu estudiat mitjançant senyals involuntaris, de manera que les respostes venen condicionades per estímuls aliens a l'àrea d'estudi. Actualment es parla d'Efecte Clever Hans en altres àmbits com és el cas de l'aprenentatge automàtic [13].

2.2.2

Àmbit de professionals de la salut

Des del punt de vista mèdic, l'aplicació d'eines d'explicabilitat en els algorismes d'IA és clau per assolir la confiança necessària per part dels usuaris i usuàries finals, que difícilment confiaran en un algorisme del tipus caixa negra. Els algorismes d'IA neixen per donar suport a la presa de decisions per part del personal sanitari i, en aquesta relació, és imprescindible la confiança.

En aquest sentit, els diversos tipus d'explicabilitat existents poden donar informació a diferents nivells, així com també poden presentar les conclusions en diversos formats, adaptant-se al cas d'ús i les necessitats dels experts i expertes. D'aquesta forma, una explicació de primer nivell (o global) permet entendre les característiques generals que té en compte un model determinat, proporcionant rànquings d'importància de característiques que expliquen quines variables influeixen més en les prediccions. En canvi, una explicació de segon nivell (local) permet identificar quines característiques són rellevants per a un o una pacient en concret mitjançant gràfics, imatges o valors numèrics, en funció de les necessitats. És per

això que és important que les persones que desenvolupen les solucions d'IA comptin amb l'assessorament del personal sanitari implicat durant el procés d'implementació de mètodes d'explicabilitat. El format i la tipologia de les explicacions haurien de ser consensuats, per així assolir un compromís entre allò que tècnicament és possible i allò que és útil per a les persones usuàries finals.

La XAI, doncs, facilita l'anàlisi per part de les persones usuàries finals, de manera que permet identificar de manera ràpida quines característiques tenen més pes en una predicció o quins punts d'una prova d'imatge han contribuït a un cert diagnòstic.

2.2.3

Àmbit de la ciutadania

El fet que les eines d'explicabilitat facilitin el coneixement dels mecanismes algorítmics, tant per part de desenvolupadors i desenvolupadores d'IA com de professionals de la salut, fa que les solucions d'IA esdevinguin més fiables. Aquest augment de la fiabilitat acaba repercutint en una major confiança per part de les persones, sobre les quals recau, en última instància, l'ús de les solucions d'IA. En definitiva, el procés de desenvolupament d'una eina d'IA ha d'estar centrat en la persona, que té el dret de conèixer com s'han pres les decisions que l'afecten.

2.2.4

Àmbit de regulació

La XAI aporta un grau de transparència a la IA, que facilita la confiança per part dels reguladors que marquen les directrius en IA. Desxifrar els algorismes de caixa negra s'ha convertit en quelcom essencial per garantir la fiabilitat de la IA i per a l'aplicació de la IA en medicina.

2.3

Preguntes que ajuden a respondre la XAI

Mitjançant l'explicabilitat podem adreçar un conjunt de preguntes obertes davant l'execució d'un algorisme d'IA. Aquestes qüestions responen a diferents àmbits:

Correcció: tenim la seguretat que totes i només les característiques d'interès han contribuït en les decisions del nostre algorisme?

Robustesa: tenim la seguretat que el model no és susceptible a pertorbacions?

Biaix: som conscients d'algun biaix específic de les dades que penalitzi injustament grups d'individus?

Millora: de quina manera concreta es pot millorar el model de predicció?

Transferibilitat: de quina manera concreta es pot aplicar el model de predicció d'un domini d'aplicació a un altre domini d'aplicació?

Comprensió humana: tenim la capacitat d'explicar la maquinària algorítmica del model a una persona experta o fins i tot a una de profana?

2.4

Models de taxonomia XAI

Tal com s'ha esmentat anteriorment, el procés de desenvolupament d'una solució de XAI ha d'involucrar especialistes que faran ús de l'eina per tal de garantir que les explicacions donades s'adeqüen a criteris mèdics i que responen a les seves necessitats pel que fa al nivell de comprensió.

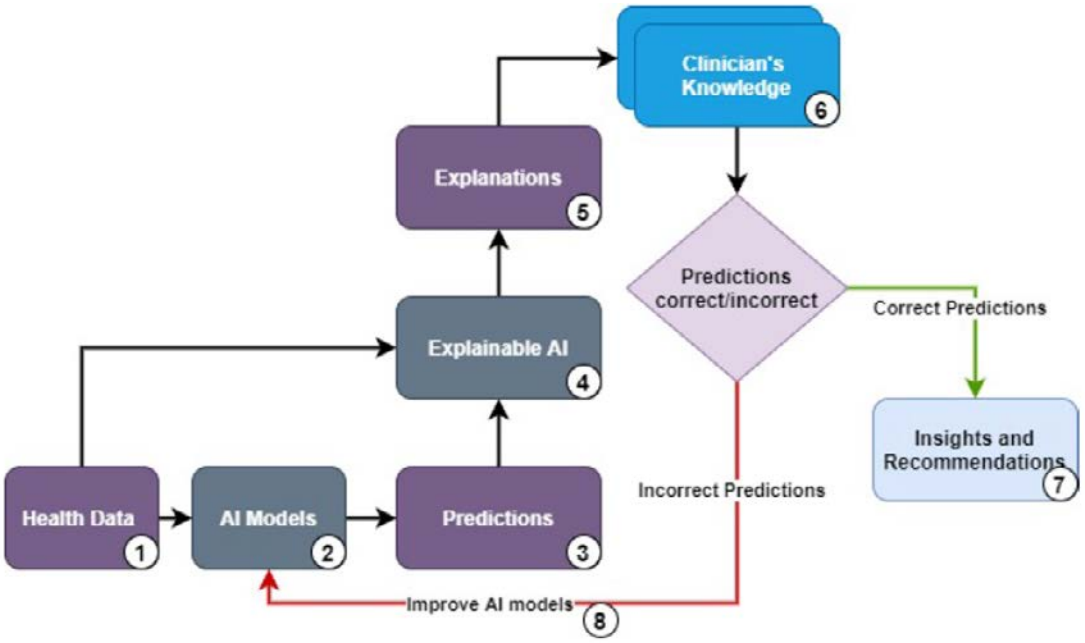


Figura 1. Esquema sobre la generació d'informació en XAI i experiència clínica [11].

- Les aplicacions de salut intel·ligents s'entrenen amb un conjunt de dades (1) i utilitzen els models resultants (2) per fer una predicció (3).
- Els models obtinguts són utilitzats per les tècniques XAI (4) per generar explicacions (5).
- Aquestes explicacions, juntament amb les prediccions, s'analitzen amb l'ajuda del coneixement de professionals de la salut especialistes en cada cas (6).
- Si les explicacions obtingudes no satisfan les persones expertes implicades, malgrat que les prediccions del model d'IA siguin encertades, cal revisar quines accions cal emprendre per millorar l'explicabilitat. En cas que es detecti que el model es fixa en característiques clarament errònies, caldrà tornar al model per analitzar possibles millores (8). També es pot donar el cas que calgui modificar la parametrització de la pròpia eina d'explicabilitat o canviar de tècnica perquè els resultats no s'ajusten al model en qüestió. Aquest procés es repetirà de forma iterativa fins a arribar a un resultat satisfactori.
- Si les prediccions i corresponents explicacions són validades per experts i expertes, podem pensar que l'algorisme es fixa en paràmetres que tenen sentit i que aquesta explicació satisfà les seves necessitats (7).

3.
Taxonomia de les eines d'explicabilitat

3.1.

Taxonomia XAI

En general, existeix una manca de consens en la classificació de les tècniques que segueixen els models XAI. A continuació es poden veure una sèrie d'exemples diferents de models de taxonomia.

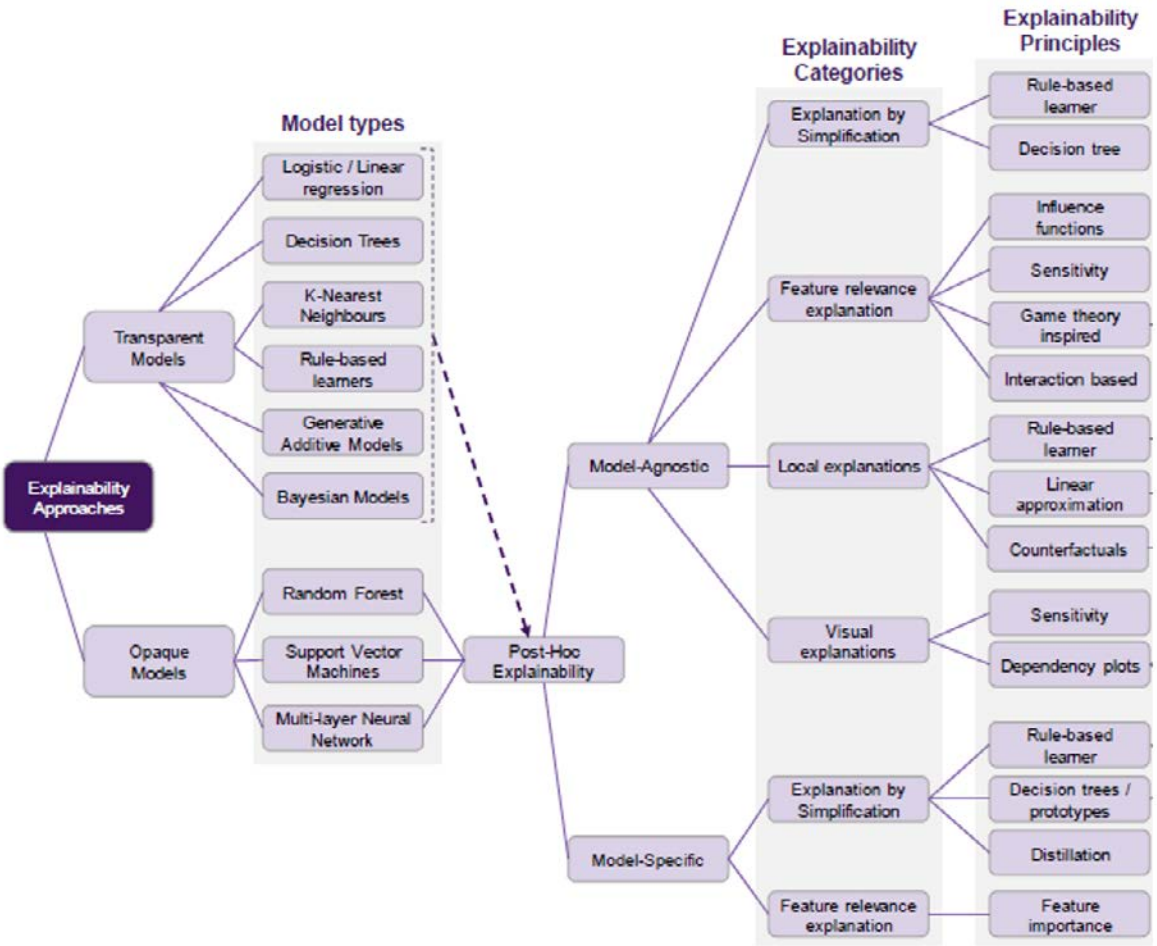


Figura 2. Exemple 1 d'un mapa de taxonomia en XAI [7].

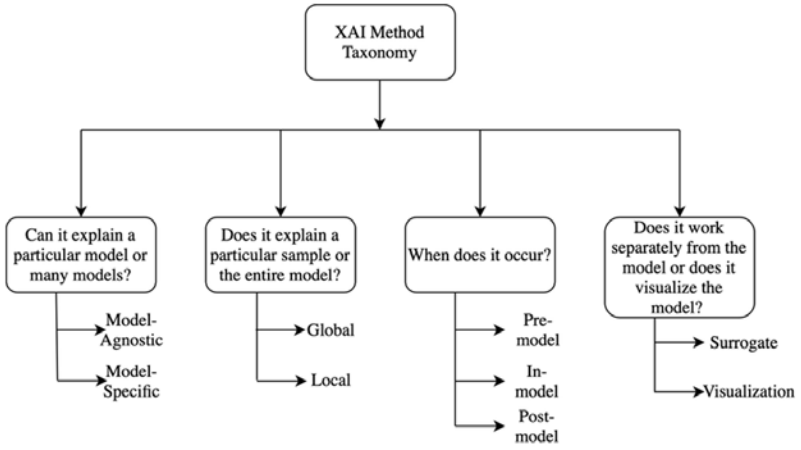


Figura 3. Exemple 2 d'un mapa de taxonomia en XAI [8].

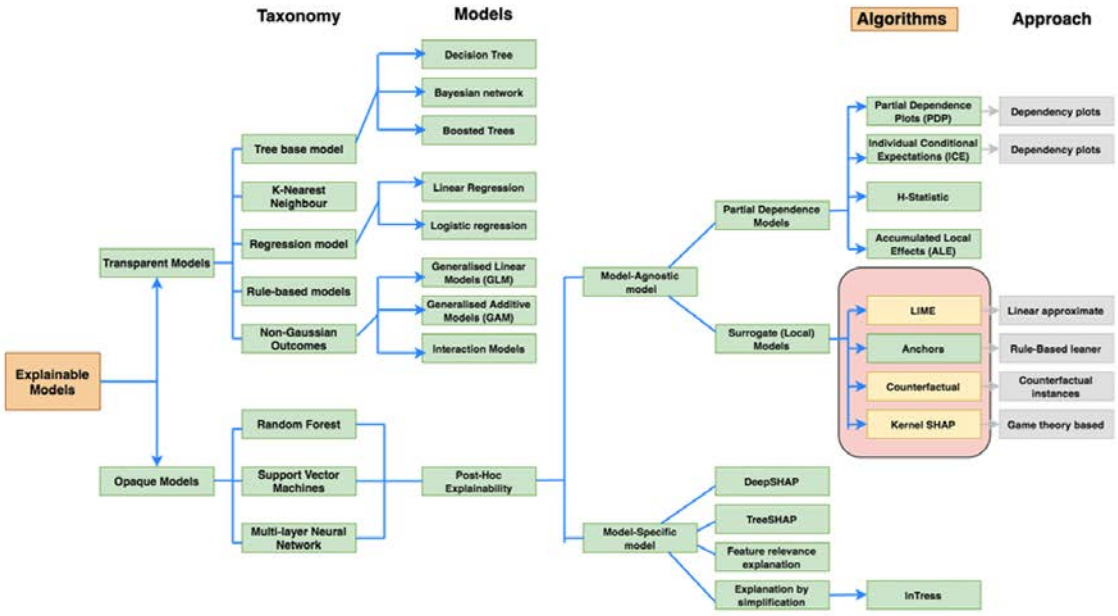


Figura 4. Exemple 3 d'un mapa de taxonomia en XAI [9].

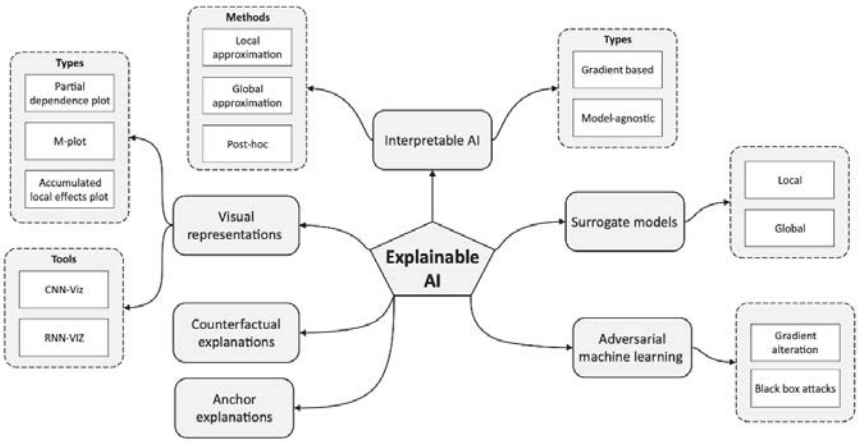


Figura 5. Exemple 4 d'un mapa de taxonomia en XAI [10].

Com es pot observar, els mètodes per a l'aprenentatge automàtic poden ser classificats segons diversos criteris.

3.2.

Explicabilitat intrínseca i explicabilitat post hoc

Distingim entre si l'explicabilitat s'obté per la naturalesa intrínseca del model o bé aplicant mètodes que analitzen els models després de l'entrenament (post hoc). **L'explicabilitat intrínseca** es refereix a l'aprenentatge automàtic que es considera interpretable per la seva simple estructura, per exemple, arbres de decisió curts o models lineals dispersos. **L'explicabilitat post hoc** es refereix a l'aplicació de mètodes d'interpretació després que el model sigui entrenat.

3.3.

Explicabilitat global i explicabilitat local

Les tècniques de XAI es poden aplicar **globalment,** mostrant una explicació general del model en el seu conjunt (importància de les característiques d'entrada en la predicció de la sortida), o bé **localment**, centrant-se en l'estudi d'un cas particular (un/a pacient específic/a).

Els mètodes globals descriuen el comportament a grans trets d'un model d'aprenentatge automàtic. Són mètodes útils quan el modelador vol interpretar i analitzar els mecanismes generals. Els mètodes locals serveixen per entendre millor les prediccions per a cada cas concret.

3.4

Models transparents i models opacs

Els models generats en IA poden ser transparents (p. ex., model obtingut d'una regressió logística) o opacs (p. ex., model obtingut d'una xarxa neuronal convolucional). La transparència representa una comprensió a nivell humà del funcionament intern del model. Es poden considerar tres dimensions. La **simulabilitat** és el primer nivell de transparència i es refereix a la capacitat d'un model de ser simulat per un humà. Només models que són simples i compactes encaixen amb aquesta categoria. En el segon nivell hi ha la **descomposabilitat**, que denota la capacitat de descompondre un model en parts (entrada, paràmetres i càlculs) i explicar aquestes parts. El tercer nivell de transparència expressa la capacitat d'entendre el procediment pel qual passa el model per tal de generar la seva sortida. Es coneix com a **transparència algorítmica** i ha de permetre inspeccionar el model amb una anàlisi matemàtica.

Els **models transparents** són un conjunt de models l'arquitectura dels quals satisfà almenys un dels tres nivells de transparència. Alguns models transparents serien [7]:

- **Regressió lineal o logística:** fa referència a una classe de models utilitzats per predir objectius continus/categòrics, respectivament, sota el supòsit que aquest objectiu és una combinació lineal de les variables predictores.
- **Arbres de decisió:** contenen un conjunt de sentències de control condicionals disposades de manera jeràrquica, en què els nodes intermedis representen decisions i poden ser etiquetes de classe (per a problemes de classificació) o quantitats contínues (per a problemes de regressió). Els arbres de decisió són més utilitzats quan és necessària la comprensió de l'aplicació.
- **Algorismes k-nearest neighbors:** tracten els problemes de classificació predient la classe d'un nou punt de dades tot inspeccionant les classes dels seus veïns K més propers (on la relació de veïnatge és induïda per una mesura de distància entre els punts de les dades). Aleshores s'assigna la classe majoritària a la instància en qüestió.
- **Aprenentatge basat en regles:** es construeix sobre una base intuïtiva de produir normes per descriure com el model genera la sortida.
- **Models additius generalitzats (GAMs):** és una classe de model lineal en què el resultat és una combinació lineal d'algunes funcions de les característiques de l'entrada.
- **Xarxes bayesianes:** les relacions probabilístiques entre variables es representen explícitament mitjançant un gràfic dirigit. A causa de la clara caracterització de la connexió entre les variables, examinen només les relacions probabilístiques. S'han utilitzat en una àmplia gamma d'aplicacions.

Els **models opacs** dificulten l'observació dels seus mecanismes interns. Per entendre aquests models opacs podem utilitzar diversos mètodes [7]:

- **Random Forest (RF):** es veu com una manera de millorar la precisió dels arbres de decisió on, en molts casos, pateixen de sobreajustament i conseqüentment de poca generalització. Aquest mètode combina múltiples arbres amb la finalitat de reduir el model portant-lo cap a una millor generalització de la resolució. Un bosc sencer és més complex d'explicar, en comparació amb un arbre de decisions, per la qual cosa força a aplicar una tècnica d'explicació post hoc per guanyar comprensió.

- **Support Vector Machine (SVM):** formen una classe de models profundament arrelats amb enfocament geomètric. Inicialment introduïts per la classificació lineal, es van estendre posteriorment al cas no lineal fent-los adequats per a aplicacions de la vida real. En els SVMs es troba el marge màxim entre punts de dades.
- **Xarxes neuronals artificials (ANN):** són una classe de models que s'han utilitzat àmpliament en multitud d'aplicacions. La seva comprensió matemàtica/teòrica no ha estat prou desenvolupada, cosa que els converteix en models de caixa negra. Des del punt de vista tècnic, les xarxes neuronals estan formades per capes successives de nodes que connecten les característiques d'entrada amb la funció objectiu. Cada node és una capa intermèdia que recull i agrega les sortides de la capa anterior i després produeix una sortida per si sola, passant el seu valor agregat per una funció (anomenada funció d'activació). Aquest procés continua capa per capa fins a arribar a la de sortida. Per tant, com més capes té el model més difícil es fa la seva interpretació. Algunes d'aquest tipus de xarxa són les xarxes neuronals convolucionals, les xarxes neuronals recurrents o les xarxes neuronals de grafs, entre d'altres.

3.5

Tècniques agnòstiques al model i tècniques dependents del model

Una altra classificació important consisteix a diferenciar entre les tècniques que depenen del model d'IA resultant del procés d'entrenament i aquelles tècniques que són agnòstiques al model.

Tècniques agnòstiques al model: han de ser prou flexibles per no dependre de l'arquitectura intrínseca d'un model. Poden ser útils per a arquitectures no estandarditzades o models personalitzats, pels quals les tècniques específiques no s'ajusten, o bé en casos en què es disposa de diversos models amb arquitectures diferents i es vol donar una explicabilitat homogènia.

Tècniques dependents del model: faciliten el desenvolupament d'algorismes més eficients i explicacions més específiques, basades en les característiques del propi model. La principal característica és que es limiten a arquitectures concretes.

3.6

Tipus d'explicabilitat

- Les explicacions visuals tenen com a objectiu generar visualitzacions que facilitin la comprensió d'un model. Es poden aplicar tant a imatges com a dades tabulars.
- Les explicacions per rellevància de característiques intenten explicar la decisió d'un model quantificant la influència de cada variable d'entrada. Són molt útils en models que utilitzen dades tabulars.
- Les explicacions per exemple extreuen instàncies representatives del conjunt de dades de formació per tal de demostrar com funciona el model.
- Les explicacions per simplificació es refereixen a les tècniques que s'aproximen a un model opac fent-ne servir un de més simple, que és més fàcil d'interpretar.

4.

Explicabilitat d’algorismes basats en imatge mèdica digital

La imatge mèdica comprèn el conjunt de tècniques i processos utilitzats per crear imatges del cos humà, o algunes de les seves parts, amb propòsits clínics com, per exemple, diagnosticar, tractar i/o seguir una malaltia. Alguns models han demostrat una extraordinària precisió en les tasques d'anàlisi d'imatges en els darrers anys. Un problema significatiu és que aquests models de Deep Learning són algorismes de caixa negra, per tant, són intrínsecament inexplicables.

Els mètodes d'explicació intenten mostrar el raonament en casos de classificació, preferentment, construint un grau de confiança entre el sistema, l'especialista en salut i el o la pacient. La majoria dels models de classificació d'imatge utilitzen metodologies post hoc per analitzar les característiques apreses pel model. Aquestes tècniques mostren les àrees discriminatives de la imatge. Hi ha una gran diversitat d'estudis que apliquen mètodes d'explicació post hoc per al càncer de mama, pròstata, pulmó, cerebral i fetge [14].

4.1. CAM (mapeig d'activació de classe)

El mètode CAM (Class Activation Mapping) és un dels més populars per a l'explicació visual d'imatge i és capaç de localitzar en una imatge les característiques que són responsables de la classificació en un model de xarxa neuronal.

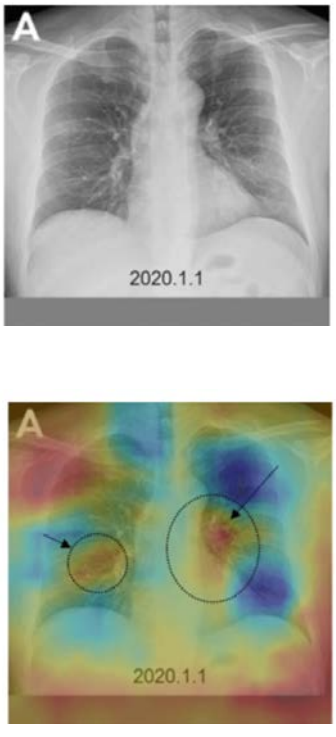


Figura 6. Representació gràfica dels resultats del mètode CAM [15].

El procés tracta d'una xarxa que consta de capes convolucionals on, just abans de la capa de sortida final, es realitza una agrupació mitjana global (GAP) en les característiques convolucionals que s'utilitzaran per a una capa connectada que produirà la sortida desitjada. Donada aquesta estructura de connectivitat senzilla, podem identificar la importància de les regions de la imatge donant un pes a la capa de sortida de les característiques convolucionals. L'agrupació mitjana global produeix la mitjana espacial del mapa de característiques de cada unitat a l'última capa convolucional. S'utilitza una suma ponderada d'aquests valors per generar la sortida final. De la mateixa manera, es calcula una suma ponderada dels mapes de característiques de l'última capa convencional per obtenir els mapes d'activació de classe [16].

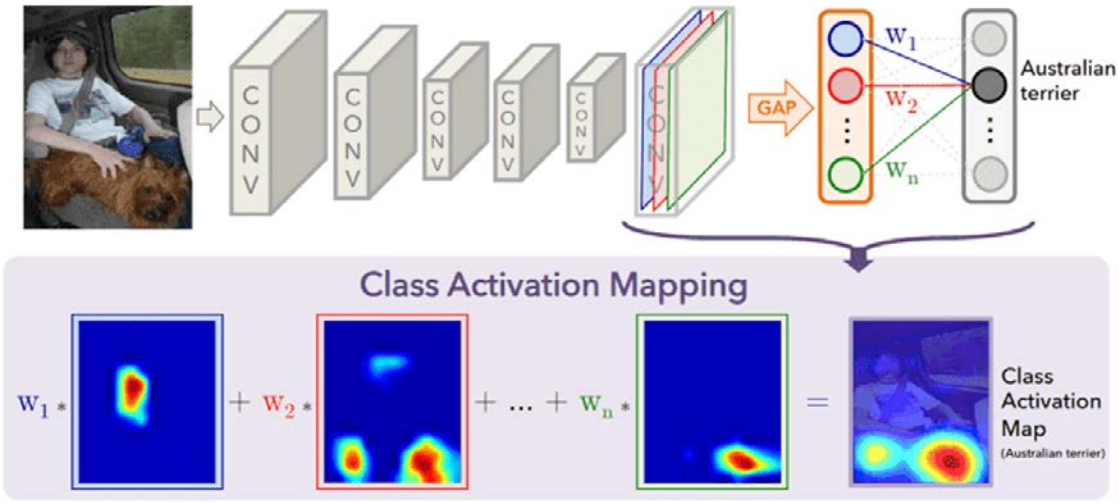


Figura 7. Procés del mètode CAM [17].

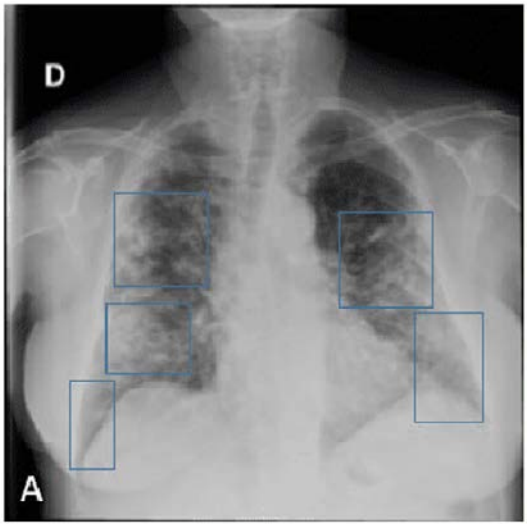
4.2.

GRAD-CAM (mapeig d'activació de classe ponderat per gradient)

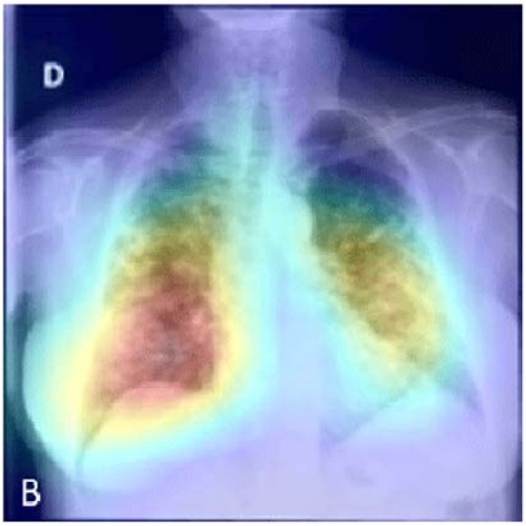
El Grad-CAM (Gradient-weighted Class Activation Mapping) és una extensió basada en el CAM, explicat anteriorment, que utilitza els gradients respecte a la classe objectiu que deriva a la capa final convolucional. El Grad-CAM produeix un mapa de localització que destaca els píxels importants per a la clas-

sificació de la imatge. A diferència del CAM, aquest mètode no requereix cap reentrenament i és àmpliament aplicable a qualsevol arquitectura basada en xarxes neuronals convolucionals (en anglès, Convolutional Neural Networks o CNN).

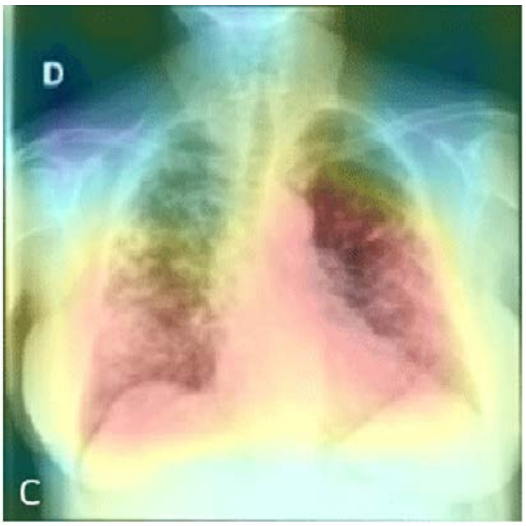
Primerament, el gradient de la puntuació de la classe es calcula respecte als mapes d'activació de l'última capa convolucional. Els gradients retornen després de fer la seva mitjana sobre la mida del mapa d'activació i, a continuació, es calculen els pesos d'importància. El factor de ponderació mostra la importància de les característiques per la classe [16].



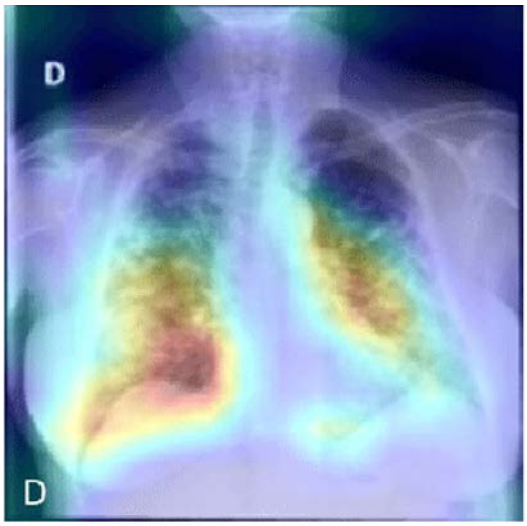
A



B



C



D

Figura 8. Visualitzacions Grad-CAM per diferents models.

4.3.

LRP (propagació de la rellevància per capes)

El LRP (Layer-wise Relevance Propagation) és una altra tècnica d'explicació visual que mostra un mapa de calor a l'espai d'entrada que indica la importància/rellevància de cada vòxel que contribueix al resultat de la classificació final. Aquest mètode no interactua amb l'entrenament de la xarxa, de manera que es pot aplicar en algorismes ja entrenats.

El LRP utilitza els pesos de la xarxa i les activacions neuronals per propagar la sortida de retorn a través de la xarxa fins a la capa d'entrada. Allà es podran visualitzar quins píxels van contribuir realment a la sortida.

La xarxa és un classificador en què cada entrada correspon a una classe diferent. En la capa de sortida, s'escull una neurona o classe que volem explicar. Per a aquesta neurona la rellevància és igual a la seva activació, per tant, la rellevància de les altres neurones en la capa de sortida serà zero. Es diu que és una tècnica conservadora, cosa que significa que la magnitud de la sortida es conserva a través del procés de retropropagació i és igual a la suma del mapa de rellevància de la capa d'entrada [19].

En èpsilon-LRP s'afegeix un petit èpsilon per propagar la rellevància amb estabilitat numèrica.

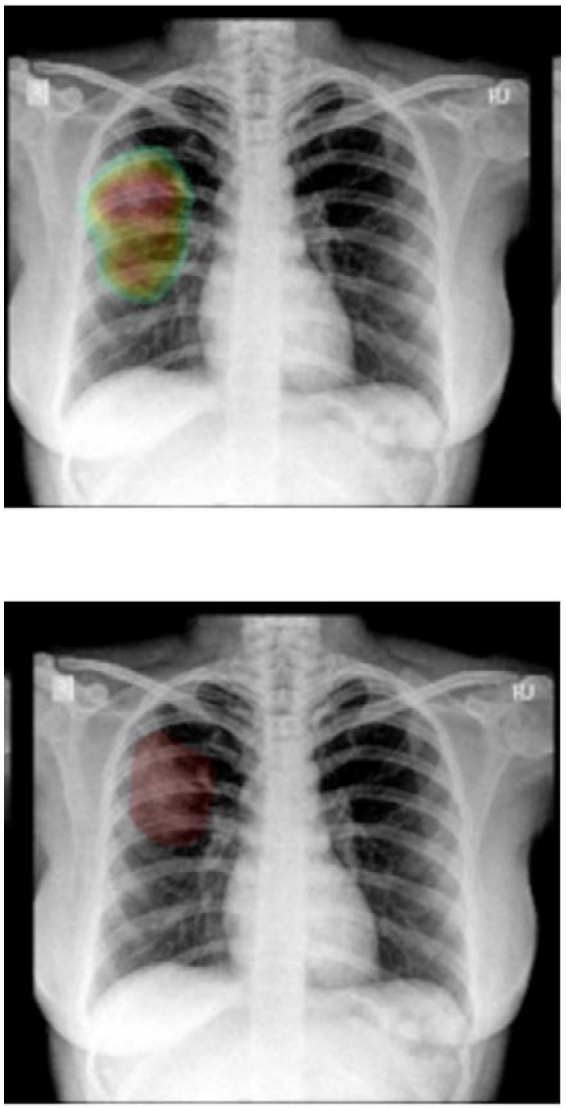


Figura 9. Mapa de calor d'èpsilon-LRP vs. l'anotació original [18].

4.4.

LIME (explicacions agnòstiques al model interpretables a nivell local)

Les LIME (Local Interpretable Model-agnostic Explanations) són explicacions que destaquen les característiques més rellevants per a la sortida. Es tracta d'un tipus d'explicació local, de manera que no intenta explicar totes les decisions que pot prendre una xarxa a través de totes les entrades possibles, sinó que només té en compte els factors que utilitza per determinar la seva classificació en una predicció individual [20].

Aquesta tècnica genera diverses mostres que són similars a la imatge d'entrada activant i desactivant alguns dels superpíxels de la imatge. El pes de cada imatge artificial per mesurar-ne la importància es calcula per explicar les característiques més importants [16].

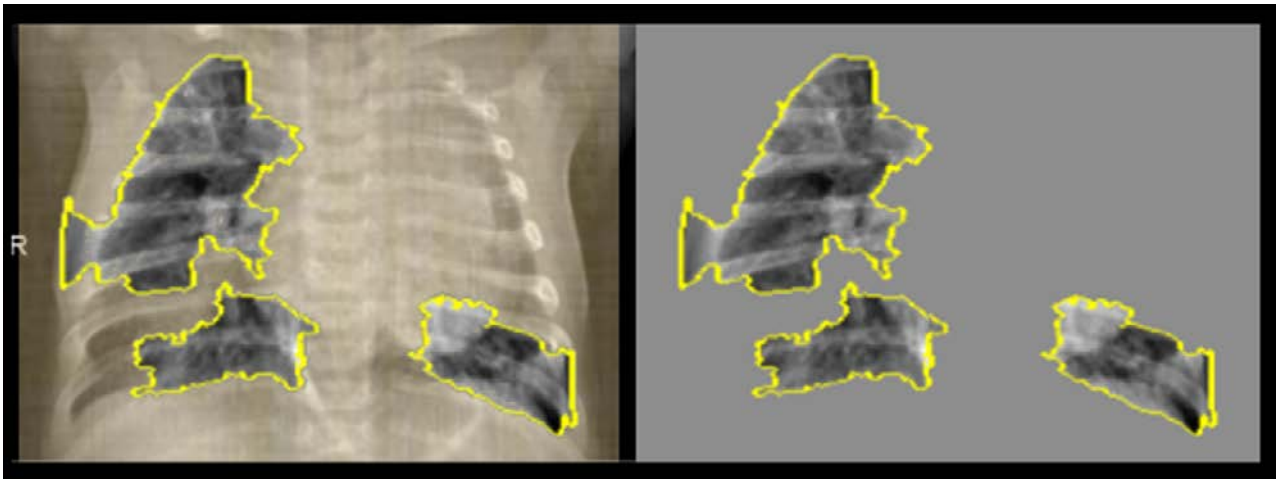


Figura 10. Mapa de calor d'èpsilon-LRP vs. l'anotació original [18].

4.5.

SHAP (explicacions additives de Shapley)

El SHAP (Shapley Additive Explanations) és un mètode per explicar prediccions individuals que es basa en els valors teòricament òptims del joc Shapley. L'objectiu és explicar la predicció d'una instància calculant la contribució de cada característica.

Els valors de Shapley sorgeixen del context on n jugadors o jugadores participen col·lectivament obtenint una recompensa p que es pretén distribuir equitativament en cada un dels jugadors o jugadores d'acord amb la contribució individual. En un model ML cada jugador o jugadora correspon a una característica i la recompensa és la predicció [21].

Les tasques de classificació d'imatge es poden explicar per les puntuacions de cada píxel d'una imatge prevista, que indica quant contribueix a la probabilitat de manera positiva o negativa.

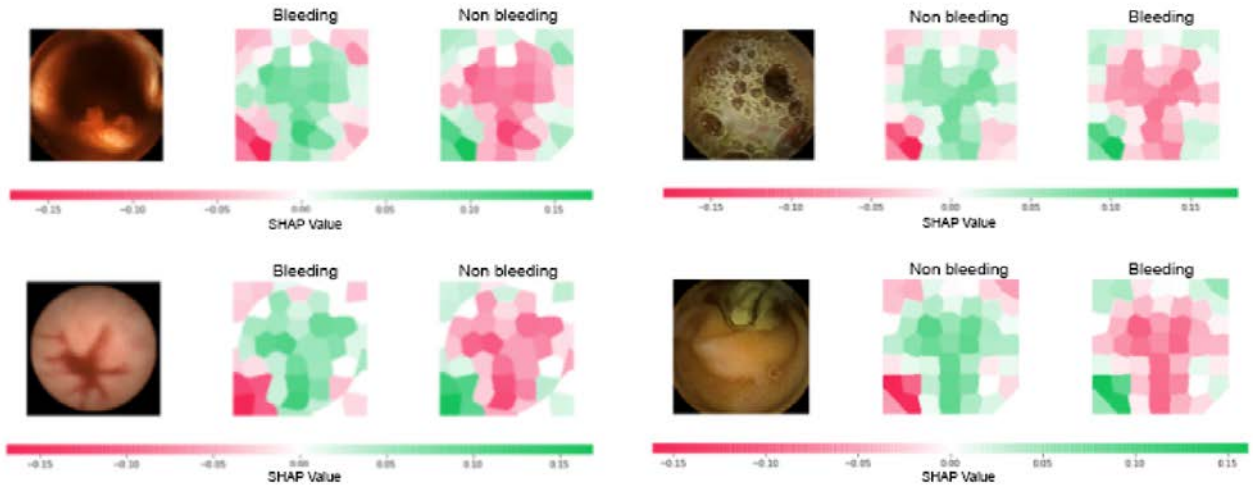


Figura 11. Explicació SHAP en imatge mèdica [22].

5.

Explicabilitat d’algorismes basats en dades tabulars

En l'àmbit de la salut s'utilitzen diàriament un gran volum de dades de tipus tabular de múltiples fonts i formats; des de variables procedents de mesures o proves directes a pacients (analítiques, constants vitals, dades òmiques, etc.) fins a registres poblacionals o dades de gestió hospitalària.

5.1

PDP (gràfic de dependència parcial)

El PDP (Partial Dependence Plot) mostra l'efecte marginal que una o dues característiques tenen en el resultat predit d'un model d'aprenentatge automàtic. A la pràctica, el conjunt de característiques S normalment només conté una característica o un màxim de dues, ja que una característica produeix parcel·les 2D i dues característiques produeixen parcel·les 3D [23].

Com a exemple, es pot veure l'efecte que tenen variables com l'edat i el nombre d'anys amb anticonceptius hormonals en una predicció de càncer:

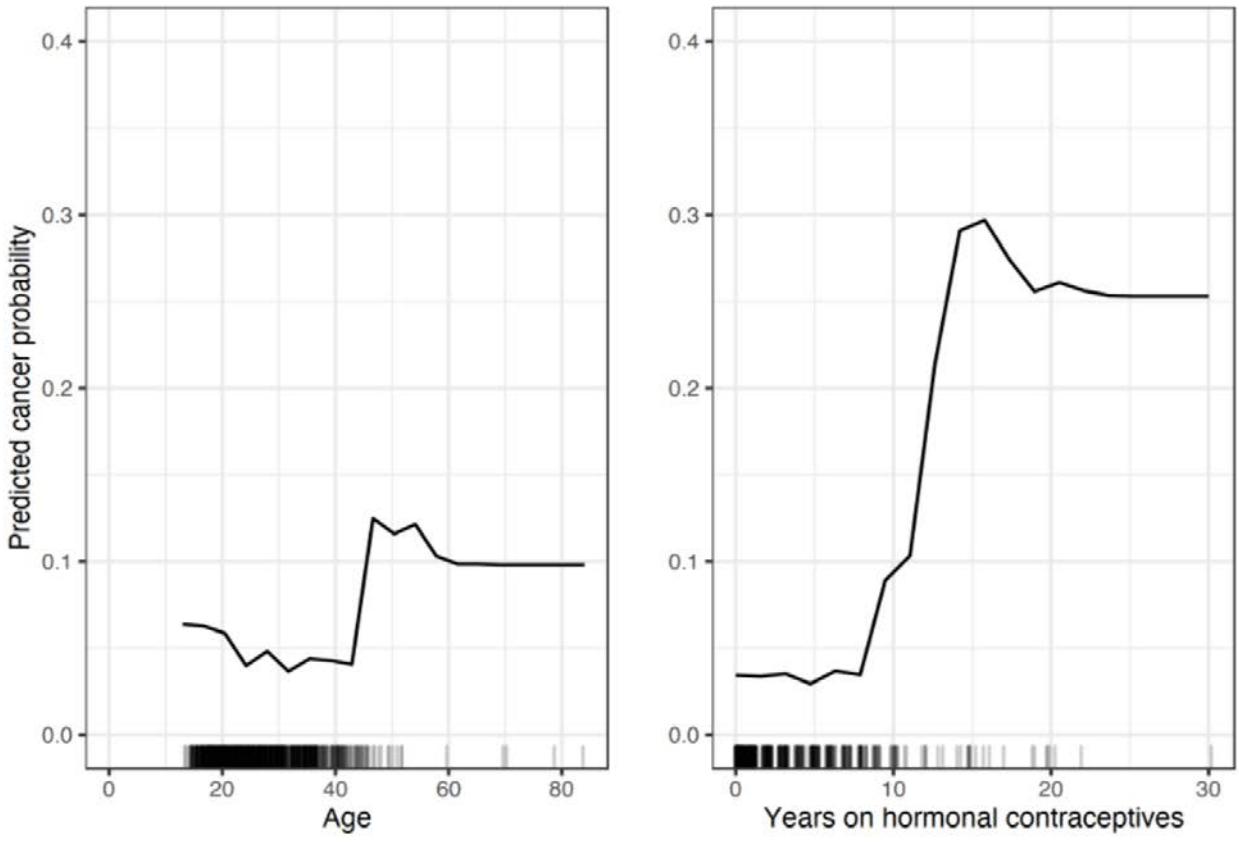


Figura 12. Explicabilitat PDP [23].

També podem visualitzar la dependència parcial de dues característiques alhora:

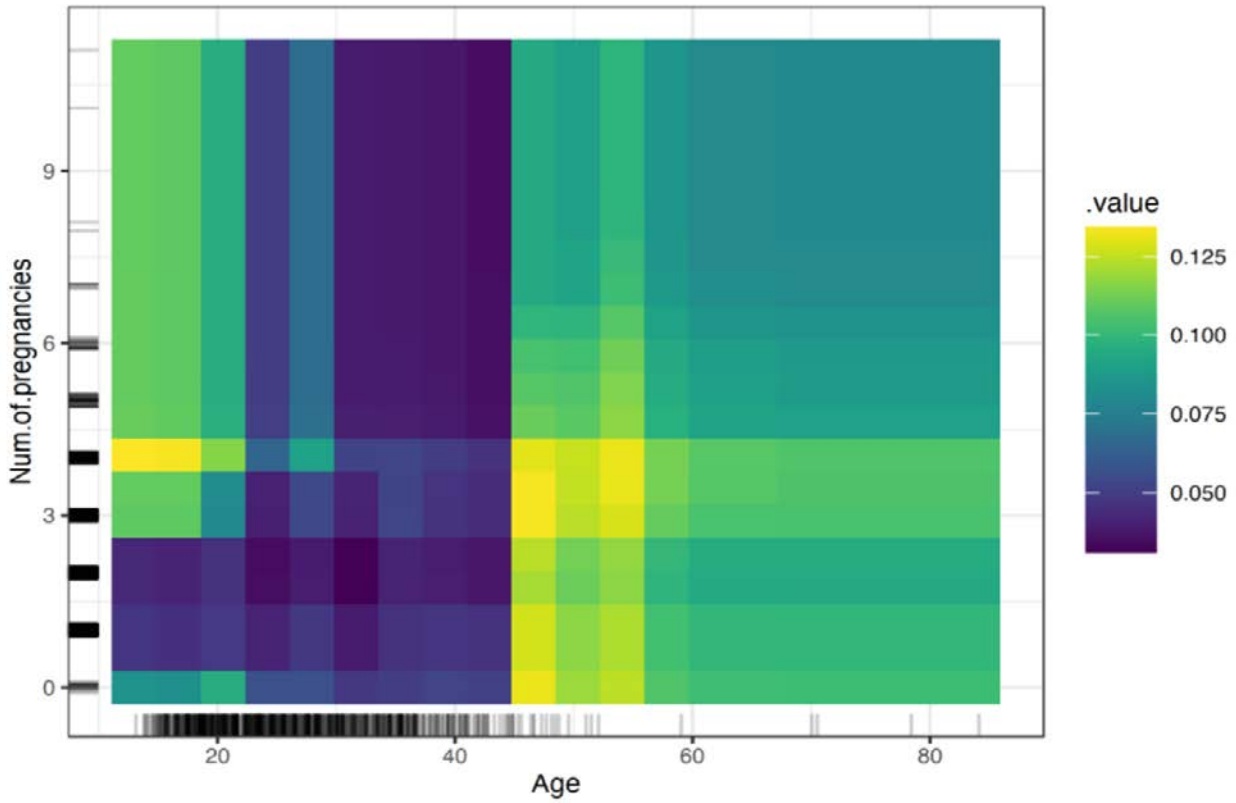


Figura 13. Representació de l'explicabilitat PDP amb 2 variables[23].

5.2

ICE (expectativa condicional individual)

Els gràfics generats per ICE (Individual Conditional Expectation) mostren com canvia la predicció de la instància quan canvia una característica. El diagrama de dependència parcial per a l'efecte mitjà d'una característica és un mètode global perquè no se centra en casos específics, sinó en una mitjana global. L'equivalent a un PDP per a instàncies de dades individuals s'anomena diagrama d'expectativa condicional individual (ICE) [24].

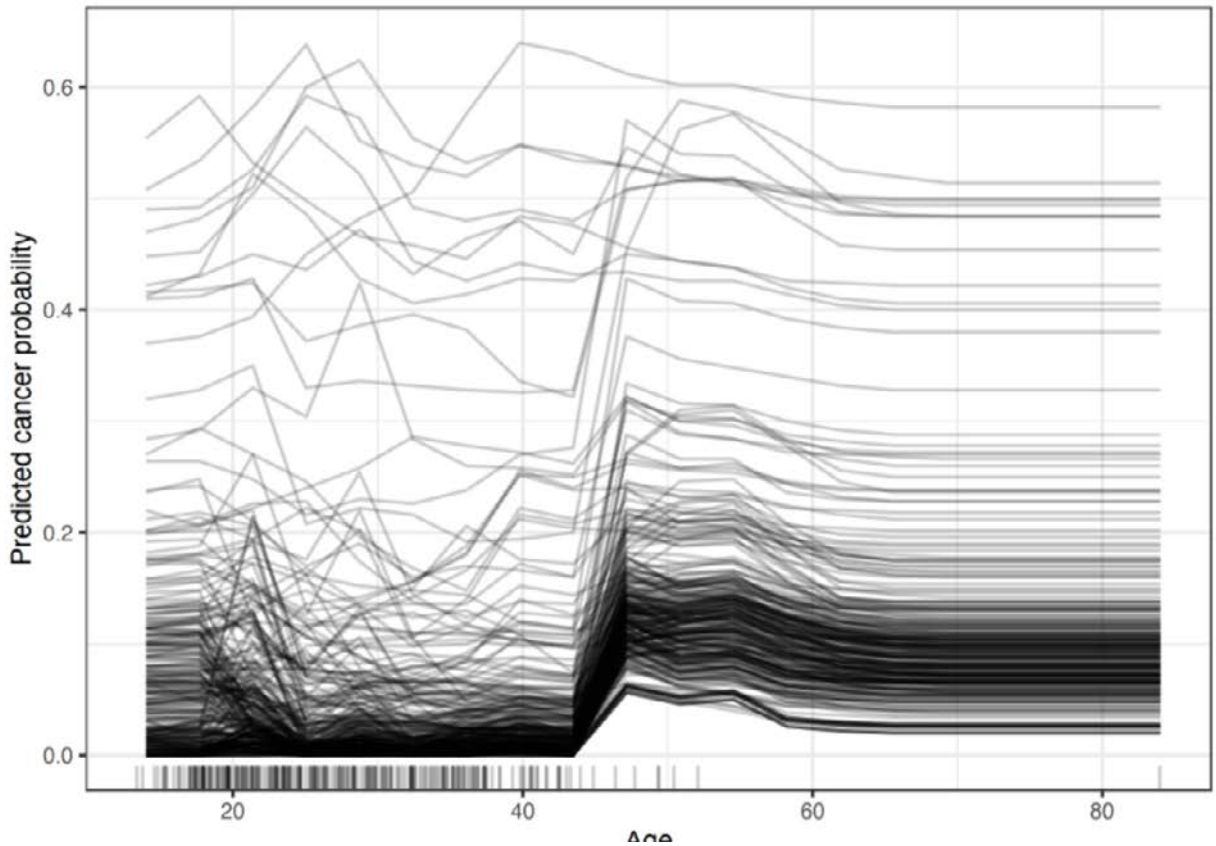


Figura 14. Explicabilitat ICE [24].

En el gràfic anterior, cada línia representa una pacient.

5.2.1

C-ICE (ICE centrat)

Hi ha un problema amb els gràfics ICE i és que a vegades pot ser difícil saber si les corbes ICE difereixen entre individus perquè comencen amb prediccions diferents. Una solució simple és centrar les corbes en un punt determinat i mostrar només la diferència fins a aquell punt. La gràfica resultant es coneix com a gràfica ICE centrada (c-ICE) [24].

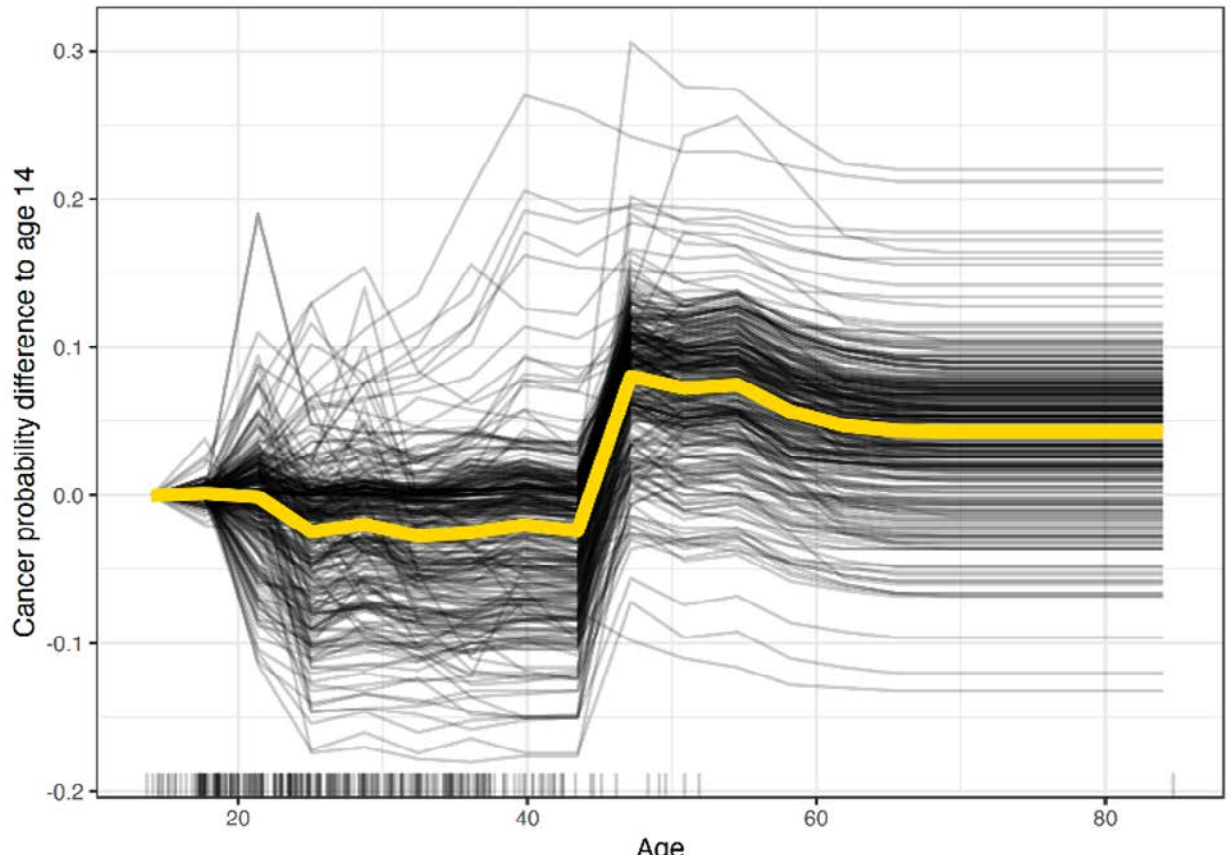


Figura 15. Explicabilitat ICE centrat [24].

5.3

Counterfactual Explanations

Aquest tipus d'explicabilitat descriu una situació causal en la forma: "Si X no s'hagués produït, Y no s'hauria produït". En l'aprenentatge automàtic explicable, les explicacions del tipus *counterfactual* es poden utilitzar per explicar prediccions d'instàncies individuals. "L'esdeveniment" és el resultat predit d'una instància, les "causes" són els valors concrets de característiques d'aquesta instància que van ser d'entrada al model i "van causar" una certa predicció [25].

Ens interessen els escenaris en què la predicció canvia de manera rellevant, com un gir en la classe predita o en què la predicció arriba a un cert llindar (per exemple, la probabilitat de càncer arriba al 10 %). Una explicació *counterfactual* d'una predicció descriu el canvi més petit en els valors de característiques que canvia la predicció d'una sortida predefinida.

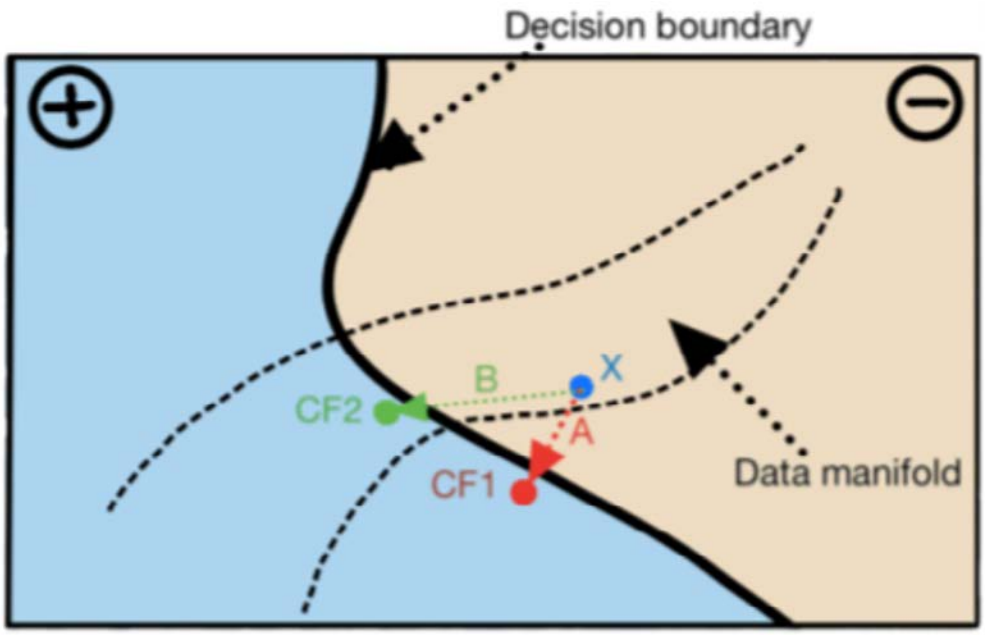


Figura 16. Explicabilitat counterfactual [25].

En el gràfic es mostren dos possibles camins per tal que un punt X (en blau), originalment classificat en la classe negativa, creui el límit de decisió. Els punts finals d'ambdós camins, CF1 i CF2, es mostren en vermell i verd respectivament.

Amb aquest tipus de mètode es pot comprovar sobre quines variables es pot incidir perquè una predicció canviï d'un estat potencialment "negatiu" a un altre de "positiu", per exemple.

5.4

LIME (explicacions agnòstiques al model interpretables a nivell local)

Podem utilitzar les LIME per a un model classificador amb imatges, com s'ha vist anteriorment, però també dades tabulars o textos.

En el cas de dades tabulars, les LIME ofereixen un tipus de gràfic explicatiu que representa la importància de cadascuna de les variables i el sentit de la seva aportació al resultat (positiu o negatiu).

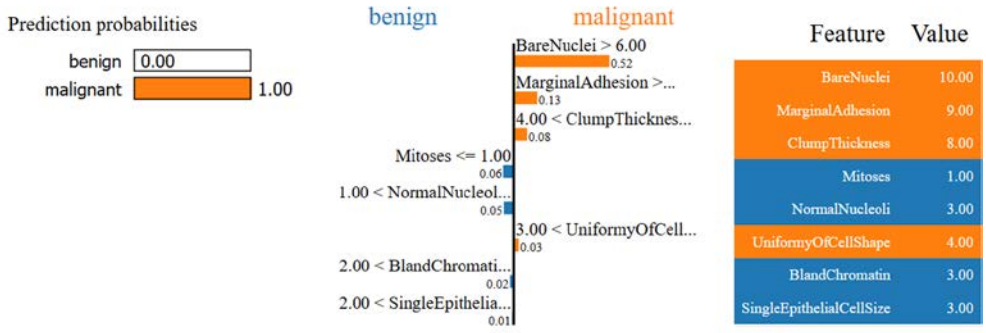


Figura 17. Explicabilitat LIME per dades tabulars.

5.5

Anchors

Aquest tipus d'explicació s'utilitza en prediccions individuals de qualsevol model de classificació de caixa negra mitjançant la recerca d'una regla de decisió que ancori prou la predicció. Una norma ancora una predicció si els canvis en altres valors de característiques no afecten la predicció.

Aquest enfocament desplega una estratègia basada en pertorbacions per generar explicacions locals per a les prediccions dels models d'aprenentatge automàtic de caixa negra i les explicacions resultants s'expressen com a regles IF-THEN fàcils d'entendre [26].

Aquest mètode proporciona una explicació de resultats com la que es mostra a continuació:

IF `PSA < 2.5 ng/ml`
AND `Age < 50`
THEN PREDICT `Cancer = false`
WITH PRECISION `97 %`
AND COVERAGE `15 %`

5.6

SHAP (explicacions additives de Shapley)

Com ja s'ha explicat abans, les explicacions de Shapley exposen la importància global de cada característica. D'aquesta manera es fa la mitjana dels valors absoluts de Shapley i s'ordenen de manera descendent segons la seva importància en la predicció final [27].

Observem l'exemple de la contribució de cada característica per a la predicció de càncer d'úter:

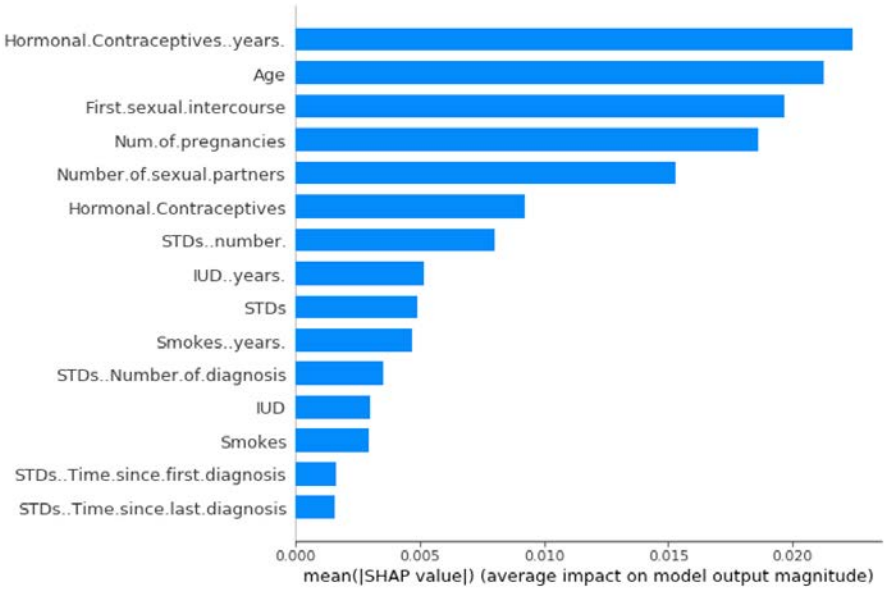


Figura 18. Gràfic d'importància de les variables amb SHAP [27].

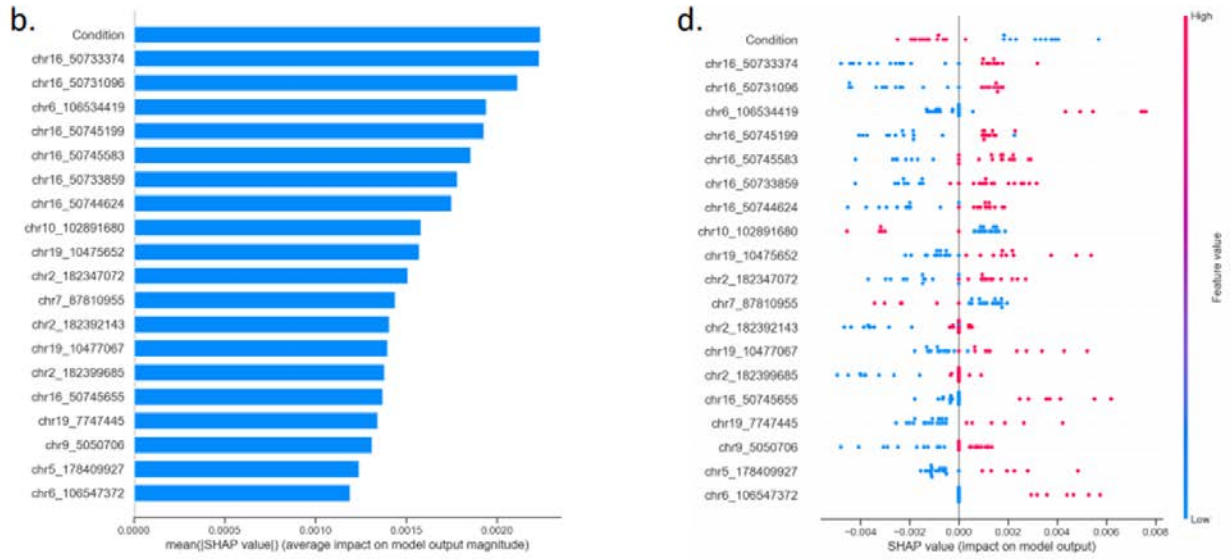


Figura 19. Gràfic d'importància de les variables amb SHAP per a la predicció amb dades òmiques

Els valors de Shapley també es poden visualitzar en forma de “forces”. Cada valor de característica és una força que augmenta o disminueix la predicció. Aquesta predicció comença en un valor basal que correspon a la mitjana de totes les prediccions. En aquest gràfic cada valor de Shapley és una fletxa que pressiona per augmentar (valor positiu) o disminuir (valor negatiu) la predicció. Aquestes forces s’equilibren entre si en la predicció real de la instància [27].

En el següent esquema (figura 26) es mostren dos gràfics d’explicació SHAP de forces per a dues pacients d’una cohort de pacients amb risc de càncer d’úter:

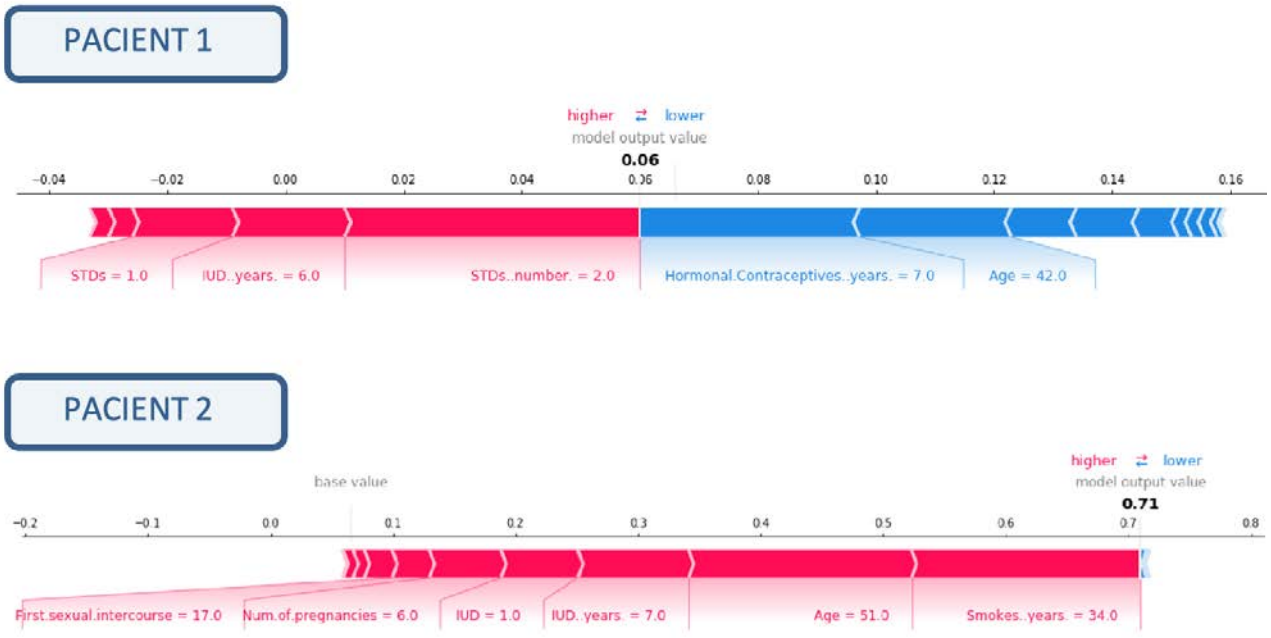


Figura 20. Diagrama de forces SHAP per a dues pacients [27].

La primera pacient té un risc del 0,06. Les variables que augmenten el risc, en vermell, es compensen amb efectes que el fan decréixer, en blau. La segona pacient té un risc més alt, del 0,71. Predominen les variables que augmenten el risc.

El SHAP proporciona múltiples formats de gràfics, que es mostren a continuació:

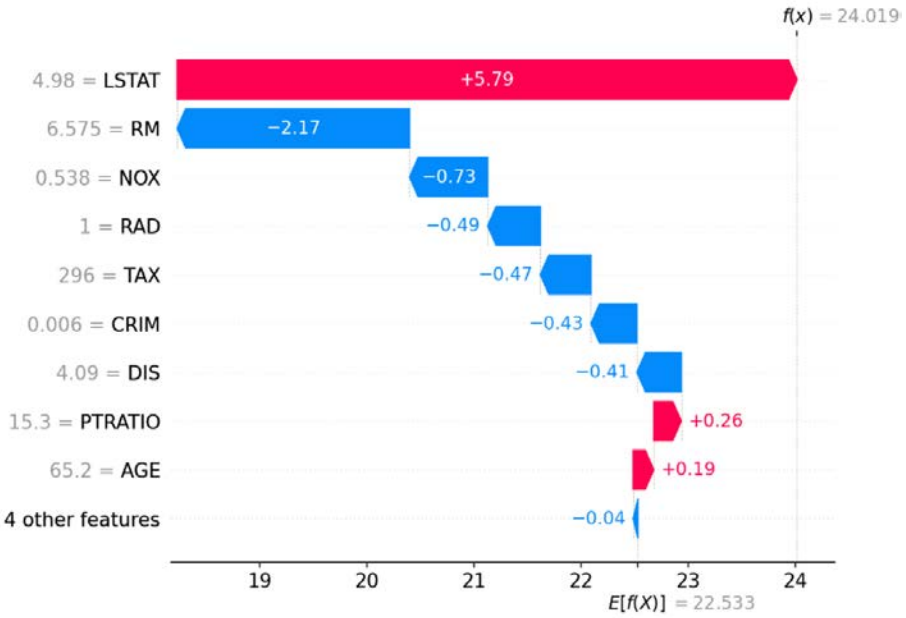


Figura 21. Representació gràfica de SHAP.

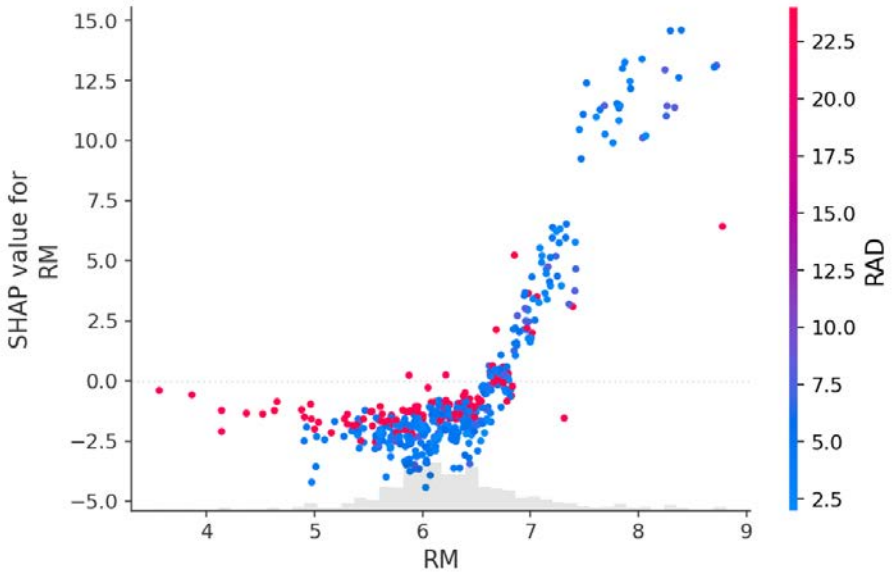


Figura 22. Representació gràfica de SHAP.

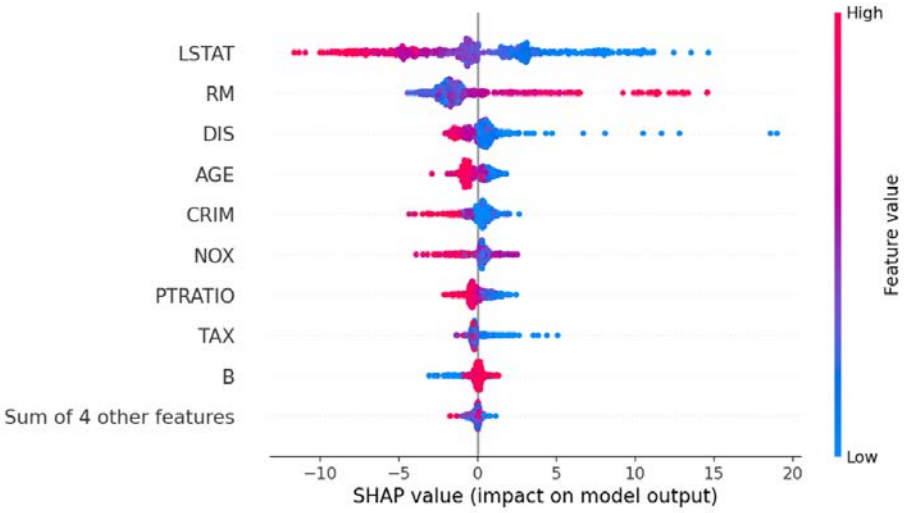


Figura 23. Representació gràfica de SHAP.

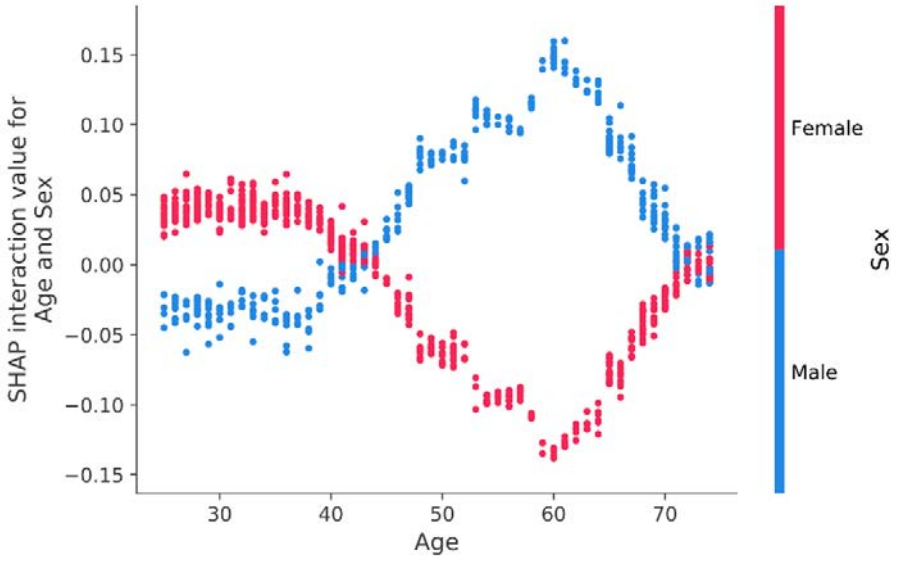


Figura 24. Representació gràfica de SHAP.

6. Explicabilitat d'algorismes basats en el processament de llenguatge natural

El processament de llenguatge natural (PLN) permet, per exemple, l'extracció d'informació estructurada a partir d'informes amb text lliure (no estructurat) amb dades de diagnòstics, tractaments o seguiment [28] [29].

Les xarxes neuronals en PLN s'entrenen d'un extrem a un altre en parelles d'entrada i sortida. Com que no es codifiquen les característiques lingüístiques explícitament, es dubta sobre quina informació es captura en les xarxes neuronals. La resposta depèn de tres elements [30]:

1. Els mètodes utilitzats per analitzar la xarxa, com de classificació o d'agrupament.
2. El tipus d'informació lingüística que se suposa que captura la xarxa com, per exemple, la longitud de l'oració, les parts del discurs o els conceptes.
3. La part de la xarxa neuronal que està sent investigada, com els pesos, activacions o incorporacions.

6.1.

Explicacions additives de Shapley (SHAP)

La tècnica SHAP també pot ser utilitzada per a l'explicació del PLN. L'objectiu en aquest cas és veure si es tracta d'informació verdadera o falsa. Aquest model ha estat prèviament entrenat amb un conjunt de dades etiquetades de manera manual. L'explicació del model és usada per explicar la sortida assignant a cada característica un valor d'importància en funció de la predicció [31].

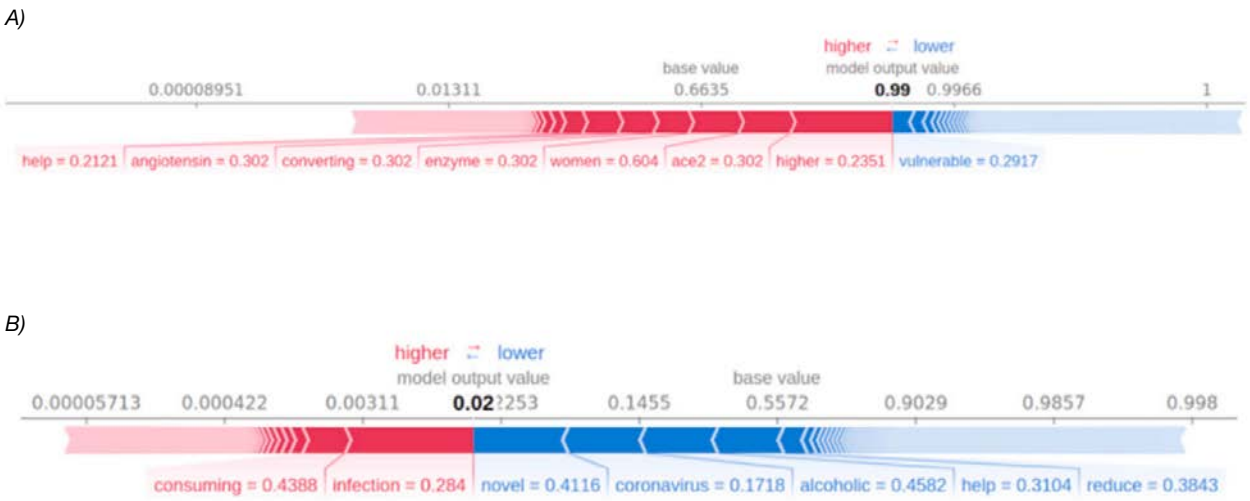


Figura 25, a. Explicabilitat SHAP en PLN. (a) Exemple d'una afirmació vertadera: "Men have higher concentrations of angiotensin-converting enzyme 2 (ACE2) in their blood than women, which may help to explain why men are more vulnerable to COVID-19 than women". I (b) d'una falsa: "Consuming alcoholic beverages may help reduce the risk of infection by the novel coronavirus". Explicades per SHAP [31].

Els factors que empenyen la predicció a ser certa es mostren en vermell mentre que els que l'empenyen a ser falsa es mostren en blau. El primer exemple (figura 25, a) representa una frase certa amb una probabilitat d'un 0,99. Les paraules que van contribuir a produir la predicció donada van ser *help, angiotensin, converting, enzyme, women, ace2 i higher.* El segon exemple (figura 25, b) representa una frase falsa amb una probabilitat de ser vertadera de 0,02. Les paraules que van contribuir a la predicció donada van ser *novel, coronavirus, alcoholic, help i reduce.*

6.2.

GbSA (anàlisi de la sensibilitat basada en gradients)

Una manera molt senzilla de relacionar les entrades amb la seva sortida és calculant la derivada parcial de la sortida respecte de cada característica d'entrada. L'anàlisi de sensibilitat es pot aplicar directament o indirectament en les dades textuals. S'obtindrà un vector de mida D amb la sensibilitat de cada sortida, on s'haurà de descompondre la norma del gradient quadrat per a la funció de predicció. Un inconvenient és que aquesta tècnica no necessàriament aplica rellevància a la característica, sinó que pot aplicar soroll.

6.3.

LRP (propagació de la rellevància per capes)

S'utilitza el LRP per descompondre la funció de decisió d'un classificador de text i utilitza les puntuacions de rellevància per proporcionar el text destacat.

A continuació es mostren uns registres en què els caràcters destacats ens ajuden a veure clarament per què el model ha predit que es tracta d'una anàlisi negativa [32].

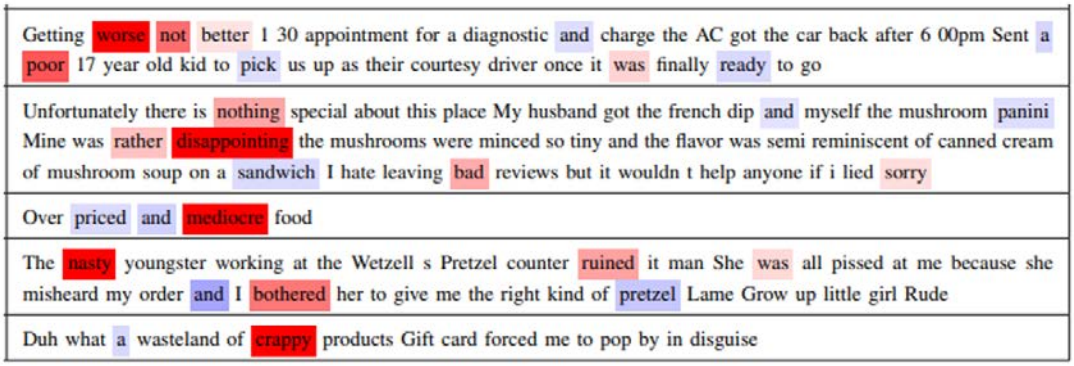


Figura 25, b. Text amb caràcters destacats segons els valors de LRP [32].

La figura 27 presenta el gràfic d'influència dels caràcters que contribueixen a una anàlisi negativa (esquerra) i a una anàlisi positiva (dreta). L'eix horitzontal representa l'impacte en el model.

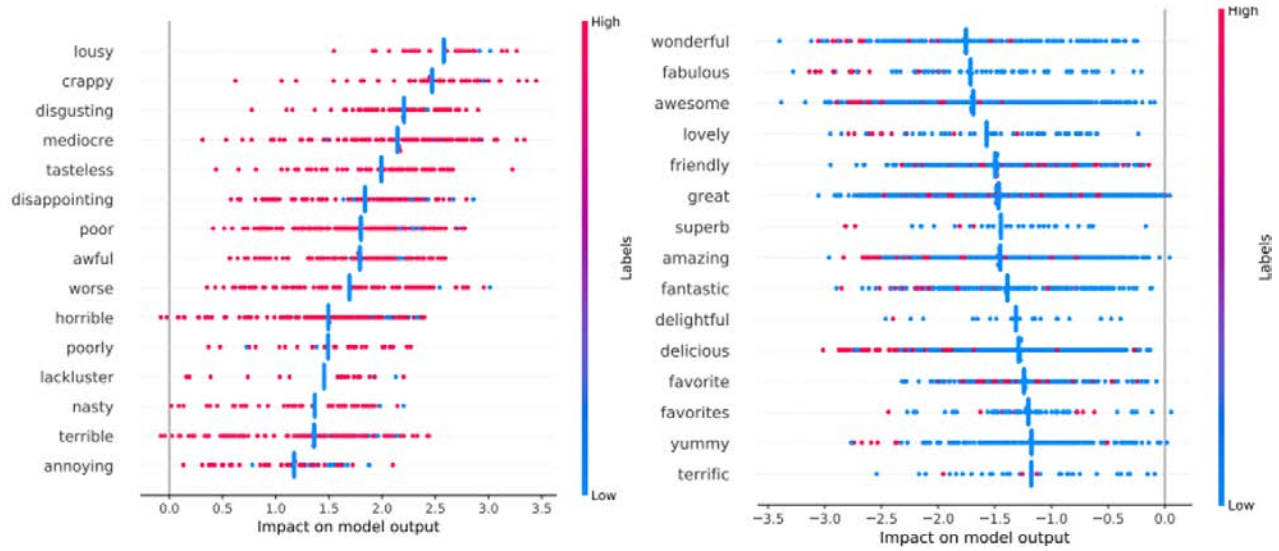


Figura 27. ***Diagrama d'unicaràcters [32].***

La figura 28 mostra la influència de la doble/triple paraula en la predicció amb els valors de LRP per una anàlisi negativa. Comparant ambdós, es pot observar que la freqüència dels caràcters és menor que en la de rellevància individual. Aquestes freqüències es poden veure a través del nombre de punts que hi ha a cada fila. La barra vertical del mig de cada fila és la contribució mitjana del caràcter sense tenir en compte si és uni, bi o tricaràcter. Per la seva baixa freqüència, els bi i tricaràcters tenen un impacte molt baix en l'anàlisi.

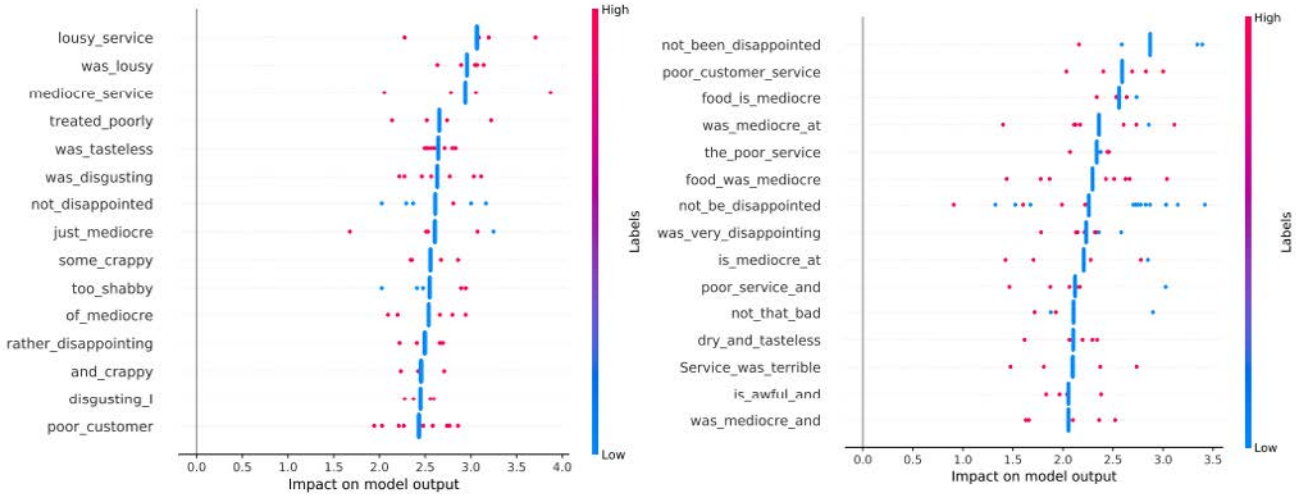


Figura 28. ***Diagrames bicaràcter i tricaràcter [32].***

6.4.

LIME (explicacions agnòstiques al model interpretables a nivell local)

Les explicacions agnòstiques al model també es poden utilitzar per entendre la sortida d'un sistema de PLN en termes humans.

Amb aquesta tècnica es va considerar que les paraules virus i protein van contribuir a definir aquest model com a virology [33].

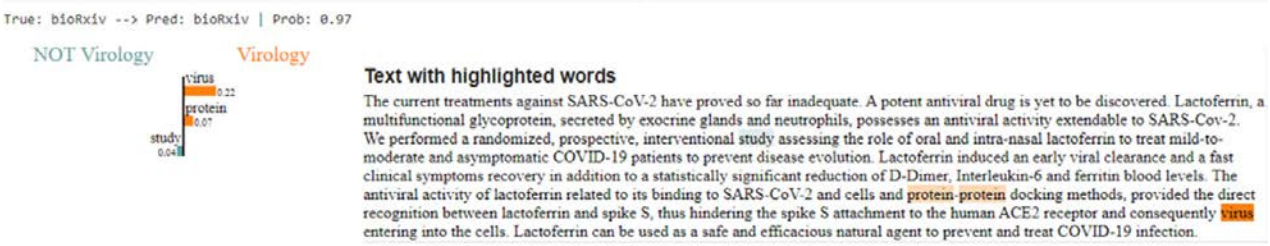


Figura 29. Explicabilitat LIME per PLN [33].

7. Referències

[1] Markus, A. F.; Kors, J. A.; Rijnbeek, P. R. “The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies”. *Journal of Biomedical Informatics,* vol. 113 (gener de 2021). Disponible en línia: https://doi.org/10.1016/j.jbi.2020.103655

[2] Maadi, M.; Khorshidi, H. A.; Aickelin, U. “A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applications”*. International Journal of Environmental Research and Public Health*, 18 (4) (22 de febrer de 2021). Disponible en línia: https://doi.org/10.3390/ijerph18042121

[3] European Comission. *High-level expert group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI*. Brussel·les: European Comission, 2019. Disponible en línia: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[4] Patel, B. N.; Rosenberg, L.; Willcox, G. [*et al.*]. “Human-machine partnership with artificial intelligence for chest radiograph diagnosis”. *npj Digital Medicine*, vol. 2, núm. 111 (18 de novembre de 2019). Disponible en línia: https://doi.org/10.1038/s41746-019-0189-7

[5] Appen. “What is Human-in-the-Loop Machine Learning?”. *Appen* [en línia] (15 de gener de 2019), https://appen.com/blog/human-in-the-loop/

[6] Vilone, G.; Longo, L. (2020). “Explainable Artificial Intelligence: a Systematic Review”. *Cornell University. arXiv* (20 de maig de 2020). Disponible en línia: https://doi.org/10.48550/arXiv.2006.00093

[7] Belle, V.; Papantonis, I. “Principles and Practice of Explainable Machine Learning”. *Frontiers* (1 de juliol de 2021). Disponible en línia: https://doi.org/10.3389/fdata.2021.688969

[8] Singh, A.; Sengupta, S.; Lakshminarayanan, V. “Explainable deep learning models in medical image analysis”*. Journal of Imaging*, 6 (6) (20 de juny de 2020). Disponible en línia: https://doi.org/10.3390/jimaging6060052

[9] Chou, Y.; Moreira, C.; Bruza, P. [*et al.*]. “Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications”*. Information Fusion*, vol. 81, p. 59-83 (maig de 2022). Disponible en línia: https://doi.org/10.1016/j.inffus.2021.11.003

[10] Nassar, M.; Salah, K.; Rehman, M. H.; Svetinovic, D. (2020). “Blockchain for explainable and trustworthy artificial intelligence”. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, vol. 10, núm. 1 (17 d’octubre de 2019). Disponible en línia: https://doi.org/10.1002/widm.1340

[11] Pawar, U.; O’Shea, D.; Rea, S.; O’Reilly, R. “Explainable AI in Healthcare”. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA)*, 2020, p. 1-2. Disponible en línia: https://doi.org/10.1109/CYBERSA49311.2020.9139655

[12] Amann, J.; Blasimme, A.; Vayena, E. [*et al.*]. (2020). “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective”. *BMC Medical Informatics and Decision Making*, vol. 20, núm. 310 (30 de novembre de 2020). Disponible en línia: https://doi.org/10.1186/S12911-020-01332-6

[13] Lapuschkin, S.; Wäldchen, S.; Binder, A. [*et al.*]. (2019). "Unmasking Clever Hans predictors and assessing what machines really learn". *Nat Commun*, vol. 10, núm. 1096 (11 de març de 2019). Disponible en línia: https://doi.org/10.1038/s41467-019-08987-4

[14] Gulum, M. A.; Trombley, C. M.; Kantardzic, M.; Martínez, I. (2021). "A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging". *Applied Sciences*, 11 (10) (17 de maig de 2021). Disponible en línia: https://doi.org/10.3390/app11104573

[15] Dialameh, M.; Hamzeh, A.; Rahmani, H. [*et al.*]. "Screening COVID-19 Based on CT/CXR Images & Building a Publicly Available CT-scan Dataset of COVID-19". *EuropePMC preprint* [en línia] (desembre de 2020), https://www.researchgate.net/figure/Plotting-the-results-of-Class-Activation-Mapping-CAM-The-CAM-highlights-the_fig4_347966202

[16] Papastratis, I. "Explainable AI (XAI): A survey of recents methods, applications and frameworks". *The AI Summer* [en línia] (4 de març de 2021), https://theaisummer.com/xai/

[17] Zhou, B.; Khosla, A.; Lapedriza, A. "Learning Deep Features for Discriminative Localization". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), p. 2921-2929. Disponible en línia: https://doi.org/10.1109/CVPR.2016.319

[18] Jain, S.; Ghosh, R. "Visualizing Deep Learning Networks - Part II". *Qure.ai Tech Blog* [en línia] (18 de desembre de 2017), https://blog.qure.ai/notes/deep-learning-visualization-gradient-based-methods

[19] Lindwurm, E. "InDepth: Layer-Wise Relevance Propagation". *Towards Data Science* [en línia] (15 de desembre de 2019), https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea

[20] Ribeiro, M. "LIME - Local Interpretable Model-Agnostic Explanations". *Homes.cs.washington.edu.* [en línia] (2 d'abril de 2016), https://homes.cs.washington.edu/~marcotcr/blog/lime/

[21] López, F. "SHAP: Shapley Additive Explanations". *Towards Data Science*. [en línia] (12 de juliol de 2021), https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

[22] "Multi-class ResNet50 on ImageNet (TensorFlow)". *SHAP latest documentation* [en línia], https://shap.readthedocs.io/en/stable/example_notebooks/image_examples/image_classification/Multi-class%20ResNet50%20on%20ImageNet%20%28TensorFlow%29-checkpoint.html

[23] Molnar, C. "8.1 Partial Dependence Plot (PDP)". *Interpretable Machine Learning* [en línia], https://christophm.github.io/interpretable-ml-book/pdp.html

[24] Molnar, C. "9.1 Individual Conditional Expectation (ICE)". *Interpretable Machine Learning* [en línia], https://christophm.github.io/interpretable-ml-book/ice.html

[25] Molnar, C. “9.3 Counterfactual Explanations”. *Interpretable Machine Learning* [en línia], https://christophm.github.io/interpretable-ml-book/counterfactual.html

[26] Molnar, C. “9.4 Scoped Rules (Anchors)”. *Interpretable Machine Learning* [en línia], https://christophm.github.io/interpretable-ml-book/anchors.html

[27] Molnar, C. “9.6 SHAP (Shapley Additive exPlanations)”. *Interpretable Machine Learning* [en línia], https://christophm.github.io/interpretable-ml-book/shap.html

[28] Danilevsky, M.; Qian, K.; Aharonov, R., [*et al.*]. “A Survey of the State of Explainable AI for Natural Language Processing”. *Cornell University. arXiv* (1 d’octubre de 2020). Disponible en línia: https://doi.org/10.48550/arXiv.2010.00711

[29] Li, Y. “Explainability for Natural Language Processing”. *Slideshare* [en línia] (5 de desembre de 2020), https://www2.slideshare.net/YunyaoLi/explainability-for-natural-language-processing

[30] Volpato, R. “NLP meets XAI: Top 5 Trends in NLP Explainability”. *Volpato.io* [en línia] (juliol de 2019), https://volpato.io/articles/1907-nlp-xai.html

[31] Ayoub, J.; Yang, X.; Zhou, F. “Combat COVID-19 infodemic using explainable natural language processing models”. *Information Processing & Management*, vol. 58, núm. 4 (juliol de 2021). Disponible en línia: https://doi.org/10.1016/j.ipm.2021.102569

[32] Gholizadeh, S.; Zhou, N. “Model Explainability in Deep Learning Based Natural Language Processing”. *Cornell University. arXiv* (14 de juny de 2021). Disponible en línia: https://doi.org/10.48550/arXiv.2106.07410

[33] Godavarthi, D.; Sowjanya, M. “Classification of covid related articles using machine learning”. *Materials Today: Proceedings*, vol. 69 (2021). Disponible en línia: https://doi.org/10.1016/j.matpr.2021.01.480

TIC Salut Social

Programa Salut/IA

Generalitat de Catalunya