DOCUMENTACIÓ DE CASOS D'ÚS

Guia d'explicabilitat de models d'IA

CTTI

3 novembre 2025 | Versió del document 1.00

Taula de contingut:

Llista d'abreviatures

Abreviatura	Significat
IA	Intel·ligència Artificial
STA	Serveis Transversals d'Analítica
PTD	Plataforma Transversals de Dades
GDPR	Reglament General de protecció de dades
SHAP	SHapley Additive Explanations
LOCO	Leave One Covariate Out
LORE	Local Interpretable Model-agnostic Explanations Local Rule-based Explanations
PDP	Partial Dependence Plot
DNN	Xarxes Neuronals Denses
CNN	Xarxes Neuronals Convolucionals
RNN	Xarxes Neuronals Recurrents
SVMs	Support Vector Machines

HISTORIAL DEL DOCUMENT

Assumpte	Data	Comentaris	Pàgines afectades
1.00		Original	Totes

1. INTRODUCCIÓ

L'explicabilitat és un element clau per garantir una integració efectiva de la intel·ligència artificial en àmbits d'alta rellevància social. A mesura que els algorismes s'apliquen en processos de decisió que afecten directament les persones —com ara els diagnòstics mèdics o les avaluacions creditícies—, resulta essencial que aquestes decisions siguin comprensibles. La capacitat d'explicar el funcionament i els resultats dels sistemes d'IA condiciona el grau de confiança que els usuaris hi poden dipositar. En absència d'explicabilitat, es pot generar una manca de confiança envers sistemes que no ofereixen justificacions clares de les seves actuacions.

Marc Ètic i Governança

L'explicabilitat no és simplement una característica tècnica desitjable sinó un imperatiu ètic. Forma part del nucli de principis per a una IA responsable juntament amb la protecció de dades i l'ús ètic, responsable i transparent dels sistemes d'IA. És per això que les decisions automatitzades, que no poden ser explicades, difícilment poden ser considerades justes o èticament acceptables, especialment en contextos delicats.

Compliment Regulatori

El marc regulatori global està evolucionant ràpidament cap a exigències explícites d'explicabilitat. El Reglament General de Protecció de Dades (GDPR) europeu va establir un precedent amb el "dret a explicació". De manera similar, regulacions emergents com l'AI Act a Europa o normatives sectorials en salut, finances i administració pública estan consolidant requisits legals específics sobre explicabilitat.

Tanmateix, la Generalitat de Catalunya ja va aprovar l'Acord de Govern ACORD GOV/45/2024, de 27 de febrer, en el qual s'impulsa la creació de la Comissió d'intel·ligència artificial i s'estableixen mesures en matèria d'intel·ligència artificial, entre les quals es troba la definició dels principis i criteris ètics en l'ús de les dades i la intel·ligència artificial de l'Administració de la Generalitat de Catalunya i el seu sector públic, aprovats per la Comissió d'IA l'11 de desembre de 2024. Inclou la creació del Registre de sistemes d'intel·ligència artificial, en el qual els sistemes d'IA productius han d'incloure una descripció sobre la seva transparència. Aquesta descripció ha de deixar clar els riscos derivats de la protecció de dades i els valors ètics.

En definitiva, aquesta guia d'explicabilitat busca donar una resposta clara a què es pot definir com a transparència, relacionant de manera directa la transparència d'un sistema d'IA amb el seu nivell d'explicabilitat.

Detecció i Mitigació de Biaixos

Els sistemes d'IA poden perpetuar o amplificar biaixos presents en les dades d'entrenament. L'explicabilitat proporciona eines crítiques per identificar aquests biaixos abans que causin danys reals. Sense mecanismes d'explicabilitat adequats, els algorismes podrien prendre decisions discriminatòries sense que es tingui capacitat per detectar-les o corregir-les.

Facilitació del Desenvolupament Científic/Tècnic

Des d'una perspectiva tècnica, l'explicabilitat permet verificar i reproduir els resultats. Un model inexplicable però efectiu pot resoldre problemes pràctics, però difícilment contribuirà a l'avanç del coneixement si opera com una "caixa negra". Aquest punt tampoc té per què ser negatiu ja que existeixen models de caixa negra que són de gran utilitat, però s'han de tenir en compte les seves limitacions i característiques.

Desafiaments Fonamentals en Explicabilitat

1. **El Dilema Precisió-Explicabilitat**

   Hi ha una tensió inherent entre el rendiment predictiu i l'explicabilitat. Els models més precisos (com xarxes neuronals profundes) solen ser menys explicables, mentre que els models més transparents

(com arbres de decisió) poden perdre precisió. Superar aquesta aparent contradicció representa un dels majors desafiaments del camp.

2. **Adaptació Contextual d'Explicacions**

   Les explicacions s'han d'adaptar a diferents audiències, contextos i nivells de coneixement. Una explicació adequada per a un enginyer de ML pot resultar incomprensible per a un metge que utilitza un sistema de diagnòstic, i viceversa.

3. **Verificació d'Explicacions**

   És clau poder donar una estandardització de mètriques per avaluar la qualitat de les explicacions.

1.1. Objectiu i motivació

Aquesta guia té com a objectiu principal proporcionar un marc tècnic i pràctic comprensiu sobre l'explicabilitat en sistemes d'intel·ligència artificial. Específicament, pretén:

1. Consolidar el coneixement actual sobre mètodes i tècniques d'explicabilitat en un recurs estructurat i accessible.
2. Oferir un full de ruta metodològic per implementar explicabilitat en projectes d'IA des de la seva concepció fins al seu desplegament.
3. Establir criteris tècnics i millores pràctiques per avaluar i garantir l'explicabilitat de models en diferents contextos d'aplicació.
4. Connectar aspectes tècnics de l'explicabilitat amb consideracions ètiques, regulatòries i de negoci.

L'elaboració d'aquesta guia respon a diverses necessitats actuals en l'ecosistema d'IA:

1. **Pressió regulatòria creixent**: Marcs regulatoris emergents com l'AI Act europea, les directrius de la FTC als EE.UU. o les normatives sectorials en finances i salut estan elevant els requisits d'explicabilitat, creant la necessitat de recursos tècnics que facilitin el compliment.
2. **Demanda de confiança**: L'adopció generalitzada de sistemes d'IA en contextos crítics requereix mecanismes que generin confiança entre usuaris i proveïdors de serveis d'IA. L'explicabilitat és fonamental per a aquesta confiança.
3. **Complexitat tècnica creixent:** Amb l'avanç cap a models cada vegada més sofisticats (transformers, models multimodals, etc.), les tècniques tradicionals d'explicabilitat necessiten adaptar-se i evolucionar.
4. **Estandardització necessària**: Manca encara d'estàndards tècnics àmpliament acceptats, la qual cosa dificulta l'avaluació i comparació de diferents enfocaments.

1.2. Audiència

Aquesta guia està dissenyada principalment per a:

1. **Científics de dades i enginyers de ML**: Professionals que desenvolupen, implementen i mantenen models d'IA/ML i necessiten incorporar capacitats d'explicabilitat.
2. **Avaluadors tècnics**: Professionals tècnics encarregats d'auditar i verificar propietats de sistemes d'IA, incloent-hi la seva explicabilitat.
3. **Responsables tècnics i funcionals**: Persones responsables de l'adopció de tecnologies basades en IA i necessiten comprendre les implicacions tècniques de l'explicabilitat.
4. **Responsables de compliment normatiu**: Persones responsables d'auditar el correcte compliment de la legislació i normes jurídiques establertes.
5. **Gestors de projectes**: Responsables de la coordinació de projectes amb components d'IA que han de garantir una integració transparent i alineada amb les disposicions en matèria d'explicabilitat.

1.3. Definicions

Per establir un marc conceptual comú, definim els termes clau següents:

1. **Explicabilitat**: Capacitat d'un sistema d'IA per presentar els seus processos, decisions i resultats de manera comprensible per als humans, revelant el "per què" i el "com" dels seus outputs de forma significativa.
    a. **Explicabilitat tècnica**: Adreçada a desenvolupadors i científics de dades, aborda els mecanismes matemàtics i algorísmics subjacents.
    b. **Explicabilitat funcional**: Orientada a usuaris professionals, explica relacions entre variables i factors de decisió sense entrar en detalls tècnics.
2. **Interpretabilitat**: Grau en què un humà pot comprendre la causa d'una decisió o predicció. Mentre que l'explicabilitat es refereix a la capacitat activa del sistema per proporcionar explicacions, la interpretabilitat denota la capacitat passiva del model de ser comprès.
3. **Transparència**: Accessibilitat i visibilitat dels components, processos i decisions del sistema. Un sistema transparent permet examinar directament els seus mecanismes interns.
4. **Caixa negra**: Model el funcionament intern del qual és opac o massa complex per ser comprès directament per humans, requerint tècniques específiques per a la seva interpretació.
5. **Mètode explicatiu ante-hoc**: Enfocament que incorpora l'explicabilitat des del disseny del model, privilegiant arquitectures inherentment interpretables.
6. **Mètode explicatiu post-hoc**: Tècnica aplicada després de l'entrenament per interpretar o explicar models ja existents sense modificar el seu funcionament intern. Aquest tipus serà el més important i sobre el que es centra aquesta guia.
7. **Explicabilitat local:** Clarificació d'una predicció o decisió específica del model, centrada en un únic cas o instància específica.
8. **Explicabilitat global**: Representació comprensible del comportament general del model en tot el seu domini operatiu.
9. **Fidelitat explicativa**: Grau en què una explicació reflecteix amb precisió el veritable funcionament del model subjacent.

1.4. Documents relacionats

1. Acord de govern: Acord adoptat pel Govern de la Generalitat de Catalunya a dia 27 de febrer de 2024.
2. AI Act: Normativa de la Unió Europea que regula el desenvolupament, ús i comercialització de sistemes d'intel·ligència artificial.
3. Canigó PTD i STA: Espai de documentació al Canigó sobre la Plataforma Transversal de Dades (PTD) i els Serveis Transversals d'Analítica (STA) i IA.
4. Principis i criteris ètics: Principis i criteris ètics en l'ús de les dades i la intel·ligència artificial de l'Administració de la Generalitat de Catalunya i el seu sector públic, aprovats per la Comissió d'IA l'11 de desembre de 2024.

2. MARC METODOLÒGIC

A continuació es detallen els punts a tenir en compte a l'hora de presentar l'explicabilitat d'un model de Machine Learning o un sistema d'IA.

2.1. Models

En un primer moment, per abordar l'explicabilitat, es fa una classificació dels models d'acord amb la seva interpretabilitat i complexitat.

2.1.1 Models altament explicables

Entre aquests tipus de models es troben els models que són fàcilment comprensibles. S'entén per comprensible al fet que les decisions preses pel model es poden explicar de manera directa.

Existeix una àmplia varietat de models que són altament explicables, entre els quals es troben:

1. **Models regressius bàsics:** Regressió lineal, regressió logística, regressions amb penalitzacions (*Ridge* i *Lasso*), entre altres.
2. **Arbres de decisió.**
3. ***Naive Bayes*.**
4. ***Clustering*:** K-Means, K-Veïns més propers i derivats d'agrupació.
5. **Sèries temporals:**
    a. SARIMAX
    b. *Holt-Winters*
    c. *State-Space Models*

2.1.2 Models de complexitat mitjana

Aquests models requereixen tècniques addicionals per explicar els seus resultats.

1. ***Random Forest***
2. ***Gradient Boosting Machines* (GBM, *XGBoost, LightGBM, CatBoost*)**
3. ***Support Vector Machines* (SVMs)**
4. **Sèries temporals**
    a. *Prophet*

2.1.3 Models de Caixa Negra

Són altament complexos i requereixen mètodes avançats d'explicabilitat. Es poden trobar:

1. **Aprenentatge profund:**
    a. Xarxes Neuronals Denses(DNN)
    b. Xarxes Neuronals Convolucionals (CNN)
    c. Xarxes Neuronals Recurrents (RNN)
2. **Models generatius (audio, imatge, vídeo, text):**

a. *Decoder-only*
b. Models BERT
c. Models BART
d. DALL·E
e. MusicLM
f. CLIP

2.2. Tècniques

Per fer que un model sigui explicable, és important considerar els següents elements:

2.2.1 Explicabilitat Global

Com s'ha esmentat anteriorment, l'explicabilitat global fa referència al comportament general del model en tot el seu domini operatiu. Per a abordar-la, s'haurà d'aportar la següent informació i haurà d'explicar-se en detall:

1. Les dades d'entrenament utilitzades i la sortida esperada. És esperable tenir el seu format, el seu volum (files i columnes) i les metadades breument resumides.
2. Identificació de les variables més importants des d'un punt de vista funcional de negoci.
3. En funció del model, es demanarà l'aplicació d'uns mètodes concrets d'explicabilitat global. Aquesta informació ve donada al punt 3 d'aquesta guia. Aquests mètodes són els següents:
    a. ***SHapley Additive Explanations* (SHAP)** [1]**:** Aquest mètode calcula la contribució de cadascuna de les variables a una predicció específica. Per això, se li assigna a cada variable un valor que representa la seva importància en la predicció, considerant totes les possibles combinacions de variables. Així, SHAP dona una base teòrica sòlida per a interpretacions consistents i equitatives.
    b. ***Leave One Covariate Out* (LOCO):** Mètode que avalua la importància d'una variable mitjançant la comparació de prediccions amb i sense aquesta variable. Es mesura com canvia la precisió o l'error del model quan una variable específica s'elimina o se substitueix per valors aleatoris, quantificant-se així el seu impacte en prediccions individuals. Pot ser considerada tant tècnica d'explicabilitat global com local.
    c. **Descomposició funcional:** Descompondre el model en components senzilles que siguin comprensibles per entendre com aquestes contribueixen al seu entrenament.
    d. ***Partial Dependence Plot* (PDP):** Mostra com una o dues variables afecten a la predicció d'un model quan la resta es mantenen constants.
    e. ***Feature Importance*:** Mesura que quantifica la contribució de cada variable en el resultat del model per a una predicció específica. A escala local, identifica quines variables van ser més determinants en una decisió particular, que permet entendre els factors clau que van influir en aquesta predicció concreta.
    f. **Mètodes propis del model.**

2.2.2 Explicabilitat Local

En aquest punt es donen els diferents mètodes d'explicabilitat local entre els que ressalten:

1. ***SHapley Additive Explanations* (SHAP).** [1]
2. ***Leave One Covariate Out* (LOCO).**

3. ***Local Interpretable Model-agnostic Explanations*** **(LIME)** [2]**:** Aquest mètode crea un model simple i transparent (com una regressió lineal) que aproxima el comportament del model complex en l'entorn local de la predicció analitzada. Per això, LIME també genera un conjunt de dades sintètiques al voltant del punt d'interès, entrena el model simple sobre ells i utilitza aquest model per a explicar la predicció original.

4. ***Local Rule-based Explanations*** **(LORE):** Enfocament que genera explicacions basades en regles lògiques per aproximar el comportament local del model. Així, amb LORE s'extreuen regles condicionals del tipus "si-llavors" que són fàcilment comprensibles per a humans i que justifiquen tant la predicció analitzada com possibles alternatives (contrafactuals).

5. ***Counterfactual Explanations*****:** Es tracten d'explicacions que identifiquen els canvis mínims necessaris en les variables d'entrada del model per alterar la predicció del model al resultat desitjat. Amb això, es busca donar resposta a la pregunta: "què hauria de ser diferent per a obtenir un resultat distint?", proporcionant-se així informació accionable sobre com modificar variables per a canviar la predicció.

3. RELACIÓ MODELS – TÈCNIQUES

En aquest apartat es dona una relació directa de les tècniques requerides (amb una creu negra) i els mètodes recomanats (amb una creu blava) amb els diferents models explicats anteriorment a l'apartat 2.1.

En primer lloc, es dona la relació dels models i els mètodes d'explicabilitat global.

Família	Model	Descomposició funcional	PDP	Feature Importance	Mètodes propis del model
Altament explicables	Models regressius	X	X		
	Arbres de decisió	X	X		X
	Naive Bayes	X	X		
	Clustering	X			X
	Sèries temporals	X			
Complexitat mitjana	Random Forest	X	X	X	X
	Boosting Machines	X	X	X	
	Support Vector Machines	X	X		
Caixa Negra	Xarxes Neuronals	X		X	X

Taula 1: relació models – tècniques d'explicabilitat global.

En segon lloc es dona la relació dels models i els mètodes d'interpretabilitat local.

Família	Model	SHAP	LOCO	LIME	LORE	Counter-factual
Altament explicables	Models regressius	X		X		X
	Arbres de decisió	X	X	X		X
	Naive Bayes					X
	Clustering	X		X		X
	Sèries temporals			X		X

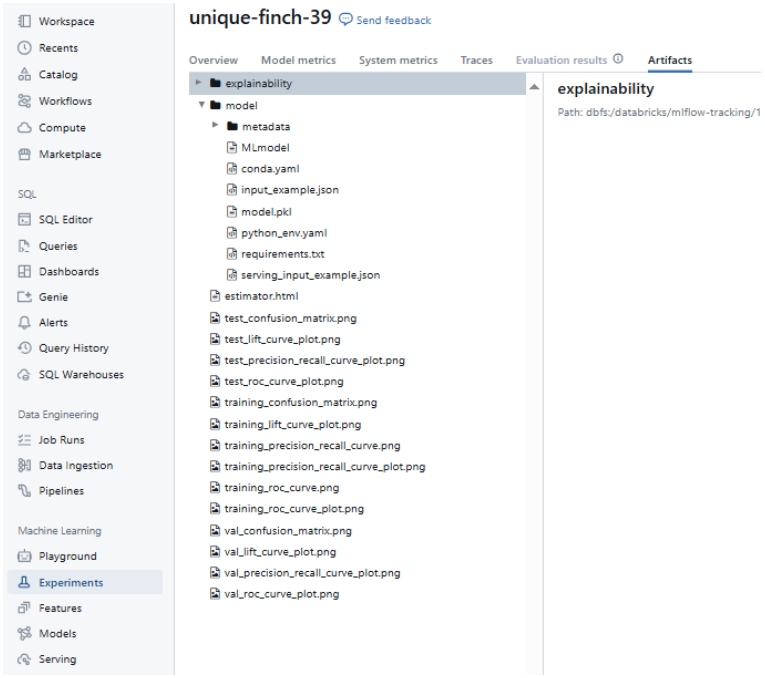
	Random Forest	X	X	X		X
Complexitat mitjana	Boosting Machines	X	X	X		X
	Support Vector Machines	X		X		X
Caixa Negra	Xarxes Neuronals	X		X	X	X

Taula 2: Relació models – tècniques d'explicabilitat local.

En aquesta relació queden exclosos els models generatius. L'explicabilitat d'aquests models ha de ser causada per la mateixa empresa desenvolupadora d'aquest. Per tant, no s'inclou en aquesta guia cap informació addicional.

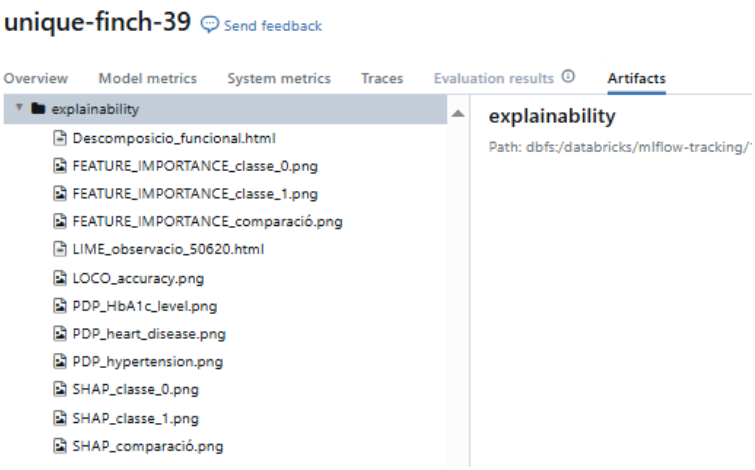
3.1. Registre de les tècniques a Databricks

Els elements de sortida de les tècniques a dalt esmentades han de quedar recollits dins dels artefactes del model, a la secció d'experiments de Databricks. Els documents generats, com ara les gràfiques, s'han de guardar dins d'una carpeta específica d'explicabilitat, tal com es pot veure en la imatge següent:



Il·lustració 2: Ubicació carpeta explicabilitat dins dels experiments

A més a més, es demana incloure noms explicatius dels gràfics generats. Aquests hauran d'incloure el nom de la tècnica utilitzada i informació rellevant, com ara la referència a l'observació que s'està analitzat o a possibles variables, entre d'altres.



Il·lustració 3: Exemple del contingut de la carpeta d'explicabilitat

4. EXEMPLES PRÀCTICS

Per tal d'aterrar el que s'ha comentat a la guia, s'ha realitzat un exemple amb un cas pràctic. El conjunt [3] de dades utilitzat és àmpliament conegut, ja que és públic i s'usa sovint en guies i explicacions diverses. Aquest conjunt està format per variables com: gènere, edat, hipertensió, malaltia cardíaca, historial de fumador/a, índex de massa corporal, nivell d'hemoglobina glicada i nivell de glucosa en sang, i s'usa per predir si un pacient té diabetis o no.

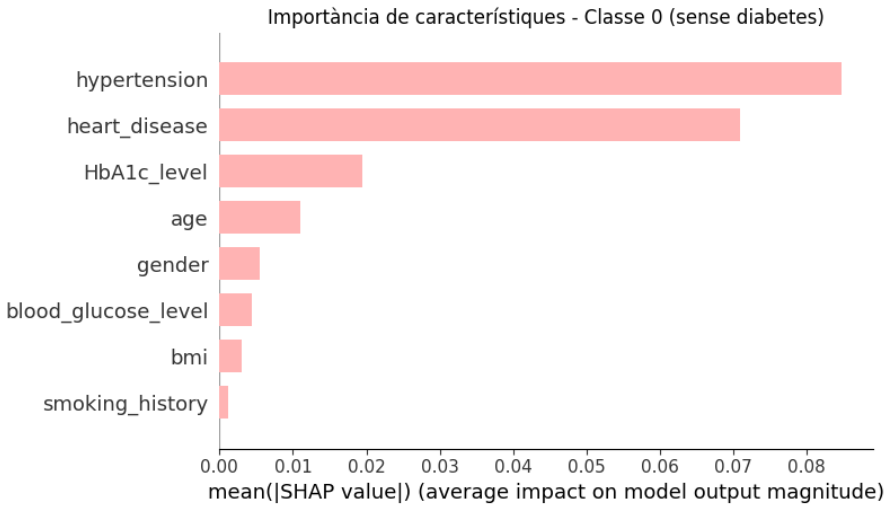
En aquest cas, s'ha usat models de *boosting* per poder classificar correctament els usuaris segons les seves característiques. Concretament, s'ha usat XGBoost per a fer-ho. Per tant, es mostraran diferents gràfics usant SHAP, LIME, LOCO, *Feature Importance, Counterfactual Explanations, Partial Dependence Plot* i descomposició funcional. No s'inclou una anàlisi detallada dels resultats, sinó que es mostren exemples de cadascuna de les tècniques d'explicabilitat.

És molt important, que sempre que sigui possible, s'usen els colors corporatius de la Generalitat. No obstant això, com es veurà a continuació, no sempre és possible. En aquests casos, s'ha de tractar d'usar colors semblants als establerts per la mateixa institució. El color corporatiu de la Generalitat és el Vermell (R: 192, G: 0, B: 0).

A més a més, sempre que sigui possible es recomana usar títols i llegendes clarificadores, en format unificat. No obstant això, tal com es mostrarà a continuació, hi ha llibreries que no donen la possibilitat de realitzar certs canvis, com ara títols, noms d'eixos o peus de pàgina.

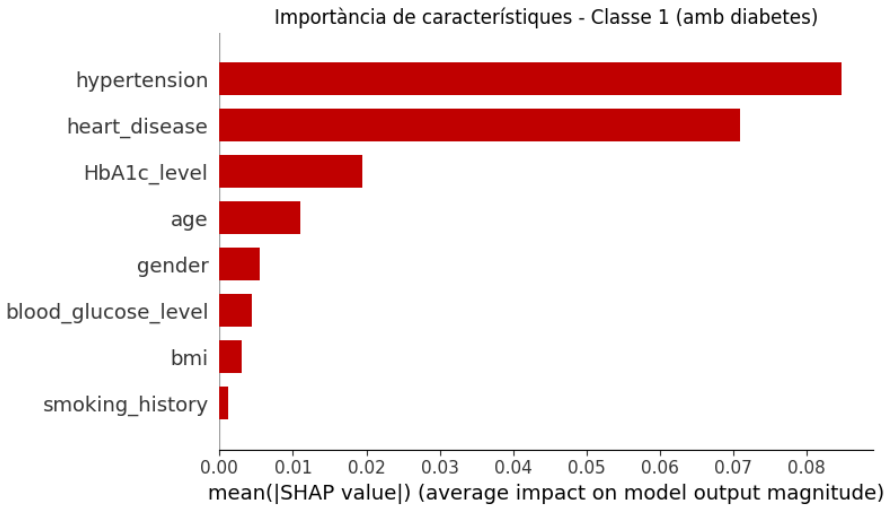
4.1. Feature Importance

Aquest primer exemple mostra la importància de les variables que s'han utilitzat per entrenar els models en funció de la classe a predir. Com que es tracta de classificació binària, en ambdós casos s'obté el mateix pes, ja que els valors SHAP per a una classe són exactament oposats als valors SHAP de l'altra classe. Això significa que la importància de les variables en termes de magnitud és idèntica per predir ambdues classes, encara que la direcció del seu impacte sigui oposada. En cas de crear-hi models de classificació amb una major varietat de categories, s'ha de realitzar aquests gràfics per a cadascuna d'elles.



Il·lustració 3: Feature Importance – Sense diabetis

El color en aquest cas és el Vermell Clar (R: 255, G: 179, B: 179).

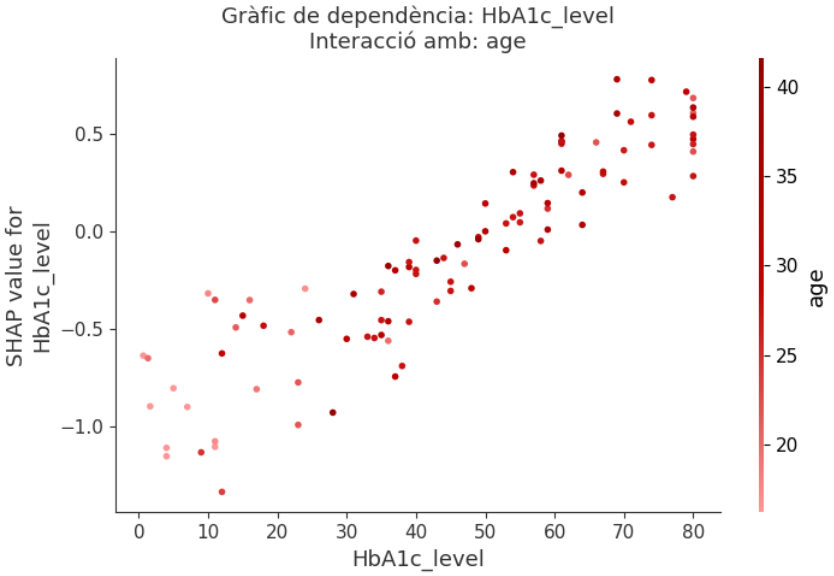


Il·lustració 4: Feature Importance – Amb diabetis

El color en aquest cas és el Vermell Corporatiu (R: 192, G: 0, B: 0).

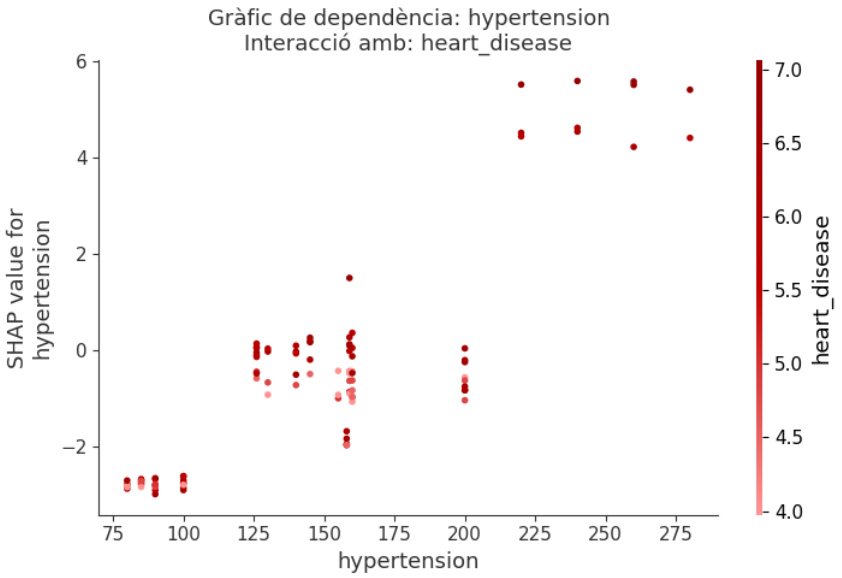
4.2. PDP

El PDP s'ha de realitzar sobre totes les variables del conjunt de dades que s'utilitza per a entrenar els models, tenint en compte, la correlació entre aquestes. És a dir, per a cadascuna de les variables que es realitza el gràfic, s'ha de calcular quina altra variable és la que té una major correlació amb aquesta.



Il·lustració 5: PDP – Variable HbA1c_level

La paleta de colors utilitzada és: Vermell Clar (R: 255, G: 179, B: 179) i Vermell Fosc (R: 128, G: 0, B: 0).

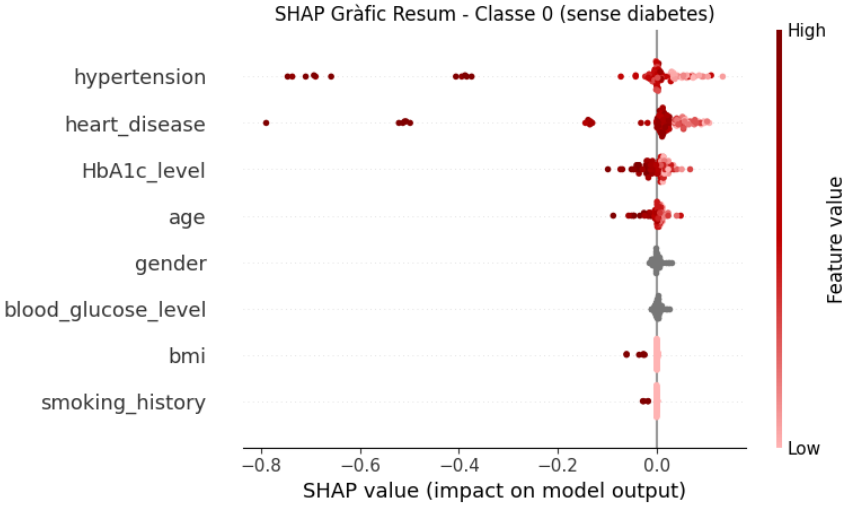


Il·lustració 6: PDP – Variable hypertension

La paleta de colors utilitzada és: Vermell Clar (R: 255, G: 179, B: 179) i Vermell Fosc (R: 128, G: 0, B: 0).

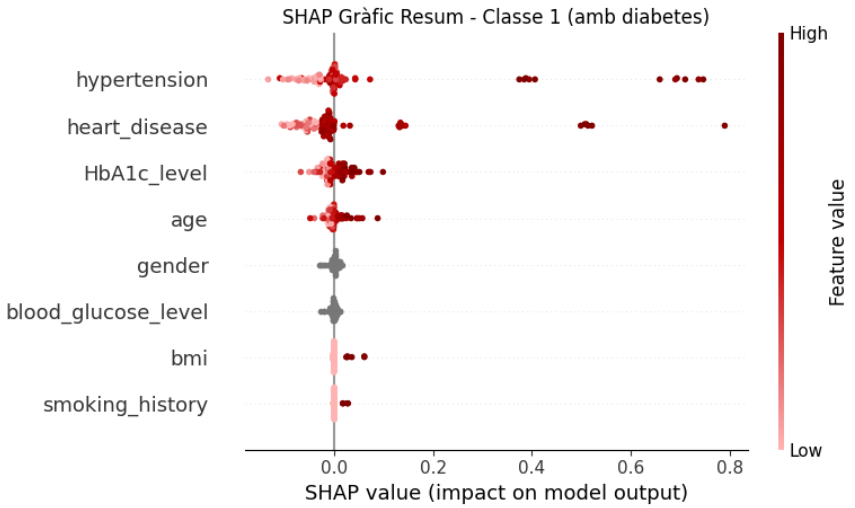
4.3. SHAP

La llibreria SHAP és àmplia i facilita la creació de gràfics molt diferents. S'exigeix reflectir els valors SHAP de cadascuna de les observacions en funció del valor de les variables que són l'input del model. De la mateixa forma que anteriorment, es realitza aquest gràfic per a cadascuna de les categories que formen part de la variable objectiva. Encara que no s'inclouen en la guia, es recomana utilitzar tants gràfics d'aquesta llibreria com sigui possible per a millorar l'explicabilitat dels models.



Il·lustració 7: SHAP – Sense diabetis

La paleta de colors utilitzada és: Vermell Clar (R: 255, G: 179, B: 179) i Vermell Fosc (R: 128, G: 0, B: 0).



Il·lustració 8: SHAP – Amb diabetis

La paleta de colors utilitzada és: Vermell Clar (R: 255, G: 179, B: 179) i Vermell Fosc (R: 128, G: 0, B: 0).

4.4. LIME

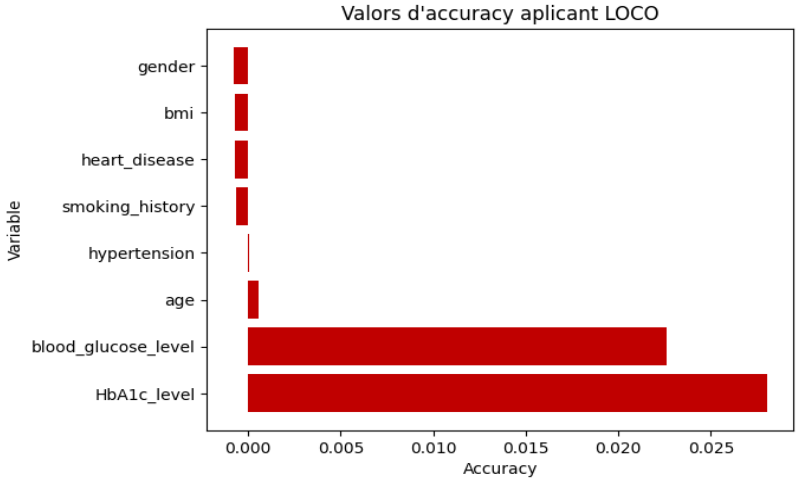
Com en SHAP, la llibreria LIME ofereix una àmplia varietat d'opcions. En aquest cas, es considera imprescindible la inclusió del gràfic següent per a diverses observacions. Com a exemple, es mostra el resultat obtingut per a un pacient, on s'avalua les probabilitats i l'impacte de les variables per obtenir aquests resultats. La quantitat d'observacions a analitzar es deixa a lliure elecció dels encarregats de desenvolupar el cas d'ús, però es recomana utilitzar diferents exemples que faciliten l'explicabilitat.

Com es pot observar, en aquest cas no és possible fer servir colors corporatius, ja que la llibreria LIME no permet modificar els colors preestablerts. Per tant, s'haurà d'usar els colors per defecte de la llibreria.

Il·lustració 9: LIME – Pacient X

4.5. LOCO

El següent gràfic mostra com varia *l'accuracy* del model en cas d'entrenar-lo amb una variable menys. Aquesta comprovació es realitza utilitzant totes les variables disponibles. Com s'observa, és possible que en alguns casos fins i tot es milloren els resultats de les mètriques obtingudes. En aquest cas, s'ha utilitzat *l'accuracy* com a mètrica per a avaluar el rendiment, però si s'utilitzen altres mètriques, s'ha de realitzar el gràfic amb les que siguin més significatives per al cas d'ús.



Il·lustració 10: LOCO – Accuracy

El color en aquest cas és el Vermell Corporatiu (R: 192, G: 0, B: 0).

4.6. Counterfactual Explanations

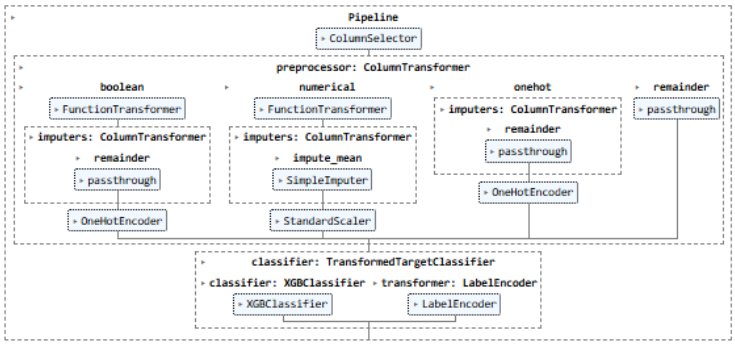
En aquest cas, no s'ha d'incloure cap gràfic necessàriament, però sí que s'ha d'analitzar diferents situacions per tal de percebre com es modifiquen els resultats obtinguts per a una observació (pacient) en funció de modificacions en les variables. Aquesta anàlisi s'ha d'incloure en el document tècnic del cas d'ús. Es mostren alguns exemples de proves realitzades sobre un dels pacients dels conjunts de dades. Aquest pacient és el mateix que l'utilitzat per a realitzar el gràfic LIME, qui té un 62% de probabilitat de no ser diabètic.

- Si el pacient, no té hipertensió, el model retornaria un 89% de probabilitat.
- Si el pacient, disminueix el bmi actual (27,32) a 22, el model retornaria un 71% de probabilitat. Per contra, si augmenta el bmi a 32, tindria un 60%.
- Si el pacient, tindria 10 anys menys, el model retornaria un 77% de probabilitat.

Addicionalment, es poden crear gràfics amb el mateix LIME, que mostren els resultats obtinguts segons les modificacions realitzades en els inputs del model.

4.7. Descomposició funcional

A continuació, es mostra la descomposició funcional del model. Aquest esquema mostra l'esquema de tots els processos realitzats, considerant tant el preprocessament de les dades introduïdes al model com l'entrenament d'aquest.



Il·lustració 11: Descomposició funcional

5. REFERÈNCIES

[1] Documentació de la llibreria SHAP. Disponible en: https://shap.readthedocs.io/en/latest/#

[2] Documentació de la llibreria LIME. Disponible en: https://lime.readthedocs.io/en/latest/usermanual.html

[3] Conjunt de dades utilitzat en l'exemple pràctic. Disponible en: https://huggingface.co/datasets/marianeft/diabetes_prediction_dataset