UNIVERSITAT DE BARCELONA

Treball de Fi de Grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Avaluació amb usuaris d'explicacions contrafactuals comparant el format tabular i el format textual

---

Gerard March Moy

Tutora: Dra. Mireia Isabel Ribera Turro

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 10 de juny de 2024

Resum

Aquest Treball Final de Grau proposa avaluar si els usuaris de peu entenen i estan satisfets amb les explicacions donades per un sistema de classificació fet amb IA, concretament amb explicacions contrafactuals aplicades a un algorisme que decideix si un préstec bancari s'atorga o no. En aquesta avaluació es realitzarà la comparació entre format textual i format tabular, per comprovar quin dels dos formats millora la comprensió i la satisfacció dels usuaris.

Per fer-ho, primer introdueix a nivell teòric el camp de la intel·ligència artificial explicable, i a nivell pràctic es prepara uns classificador binari, entrenat amb un conjunt de dades amb sol·licituds de crèdit. També es proposa el disseny experimental per a una prova d'avaluació amb usuaris. Aquesta prova consistirà en plantejar als participants una sol·licitud de crèdit que s'ha classificat com a denegada i posteriorment se'ls mostrarà els contrafactuals en un dels dos formats, per posteriorment omplir dos formularis, un de comprensió i un de satisfacció.

Resumen

Este Trabajo Final de Grado propone evaluar si los usuarios de a pie entienden y están satisfechos con las explicaciones dadas por un sistema de clasificación hecho con IA, concretamente con explicaciones contrafactuales aplicadas a un algoritmo que decide si un préstamo bancario se otorga o no. En esta evaluación se realizará la comparación entre el formato textual y formato tabular, para comprobar cual de los dos formatos mejora la comprensión y satisfacción de los usuarios.

Para realizarlo, primero se introduce a nivel teórico el campo de la inteligencia artificial explicable, y a nivel práctico se prepara un clasificador binario, entrenado con un conjunto de datos con solicitudes de crédito. También se propone el diseño experimental para una prueba de evaluación con usuarios. Esta prueba consistirá en plantear a los participantes una solicitud de crédito que se ha clasificado como denegada y posteriormente se les mostrará los contrafactuales en uno de los dos formatos, para posteriormente rellenar dos formularios, uno de comprensión y otro de satisfacción.

Abstract

This TFG proposes to evaluate whether ordinary users understand and are satisfied with explanations generated by an AI-based classification system, specifically with counterfactual explanations applied to an algorithm that decides if a loan is granted or not. In this evaluation, there will be a comparison between the textual format and the tabular format to check which of the two formats improves user understanding and satisfaction.

To accomplish this, there will be a theorical introduction about the field of explainable artificial intelligence and later a binary classificatory will be prepared, trained with a dataset of credits applications. Also, an experimental design for a user evaluation test will be proposed. This test lies in outlining to the participants a credit application classified as denied and afterwards presenting the counterfactuals in one of the two formats, after which they will fill out two questionnaires, one to evaluate understanding and the second one to evaluate satisfaction.

Agraïments

Primer de tot, m'agradaria agrair a la meva tutora, la Mireia Ribera per ajudar-me en els moments que ha estat més complicat avançar, aportant els seus coneixements i una forma de treballar que m'ha ajudat a controlar els nervis i la pressió en gran part del desenvolupament del projecte. Ha estat un gran suport amb el que sempre he pogut comptar.

També m'agradaria agrair l'ajut del Jordi Vitrià, aportant suggeriments i els seus coneixements en el sector, a part de facilitar les llibreries utilitzades en aquest projecte, simplificant molt la cerca de llibreries per a preparar el projecte pràctic.

Finalment, en aquest agraïment no puc deixar de banda a totes aquelles persones que no han col·laborat de forma directa en el projecte, però que sense ells no hauria pogut ser possible. Primer de tot, als companys de la carrera que en els moments més complicats han estat al meu costat i m'han forçat a allò que possiblement em costi més, desconnectar per poder seguir endavant. També m'agradaria agrair als companys d'activitats que realitzo en el meu temps lliure per donar-me tot el seu suport quan no aconseguia els resultats que esperava, a part de donar-me un temps de distracció molt enriquidor. Finalment, m'agradaria agrair també als meus pares per animar-me quan els ànims més dequeien.

Aquest treball ha estat possible gracies a tots ells. Moltes gràcies!

ÍNDEX

1. INTRODUCCIÓ 1

1.1. MOTIVACIÓ 2

1.2. OBJECTIUS 2

1.3. ESTRUCTURA DEL PROJECTE 2

2. PLANIFICACIÓ DEL PROJECTE 3

2.1. PRESSUPOST DEL CLASSIFICADOR 3

3. APARTAT TEÒRIC 5

3.1. QUÈ ÉS UN MODEL D'INTEL·LIGÈNCIA ARTIFICIAL EXPLICABLE? 5

3.2. INTERÈS DELS MODELS D'INTEL·LIGÈNCIA ARTIFICIAL EXPLICABLES 8

3.3. DEFINICIÓ D'EXPLICACIÓ 9

3.3.1. MOTIUS DE UNA PERSONA PER DEMANAR UNA EXPLICACIÓ 11

3.3.2. COMPLEXITAT DE UNA EXPLICACIÓ 11

3.4. TIPUS D'EXPLICACIONS 12

3.5. EXPLICACIONS CONTRAFACTUALS 12

4. APARTAT PRÀCTIC 15

4.1. CONJUNT DE DADES 15

4.1.1. DADES ORIGINALS 15

4.2. LA LLIBRERIA AIX360 I LES SEVES POSSIBILITATS 20

4.3. EINES UTILITZADES PER A LA PREPARACIÓ DEL CLASSIFICADOR 21

4.4. TRACTAMENT DE LES DADES 23

4.5. PREPARACIÓ DE L'ENTORN DE DESENVOLUPAMENT I PROCÉS D'INSTAL·LACIÓ 25

4.6. PREPARACIÓ DEL PROJECTE 27

4.6.1. MODEL DARRERE EL SISTEMA 28

4.6.2. MÈTODE PER GENERAR EXPLICACIONS 31

4.7. DEFINICIÓ DEL DISSENY EXPERIMENTAL 33

4.7.1. CRITERIS D'INCLUSIÓ I EXCLUSIÓ 33

4.7.2. RECURSOS UTILITZATS 34

4.7.3. PROCEDIMENT DE LA PROVA D'AVALUACIÓ 36

4.7.4. CÀLCUL I ANÀLISI DELS RESULTATS OBTINGUTS 37

5. CONCLUSIONS 38
5.1. TREBALL FUTUR 38
6. ANNEXOS 40
6.1. GLOSSARI 40
6.2. CONTRAFACTUALS UTILITZATS PER A LA PROVA AMB USUARIS 41
6.3. GUIÓ DE PRESENTACIÓ A L'USUARI 47
6.4. CONSENTIMENT INFORMAT PER A PARTICIPAR A LA PROVA D'AVALUACIÓ 48
6.5. FORMULARI DE COMPRENSIÓ 50
6.6. QÜESTIONARI DE SATISFACCIÓ 51
7. REFERÈNCIES 53

1. Introducció

El camp de la Intel·ligència Artificial es troba en auge, amb noves tecnologies en continu desenvolupament. Tenint en compte, a més, el seu potencial, no és d'estranyar que s'implementi poc a poc en molts dels camps quotidians, ja sigui ajudant a les persones a realitzar tasques complexes o bé, realitzant les tasques per les quals està entrenada de forma autònoma.

A part de les grans capacitats que ha demostrat tenir, especialment han demostrat ser molt extrapolables, ja que són capaces de ser útils per a tasques molt variades amb l'entrenament corresponent. Per exemple, tenim l´eina ChatGPT, que rep com a entrada de dades un text, per posteriorment retornar un text que permeti respondre al text que ha rebut a l´entrada, però també tenim eines com DALL-E que, malgrat rebre el mateix tipus de input, en comptes de generar un text, genera una o varies imatges que compleixen les mateixes condicions que l'usuari ha establert amb el text d'entrada.

Però, realment som capaços d'entendre per què el model ha retornat una determinada sortida o resultat? Donada la gran complexitat darrere els models d'intel·ligència artificial, pot resultar molt complicat entendre quin es el criteri que ha aplicat el model per a generar el resultat final.

Degut a la necessitat de poder comprendre les decisions que ha pres un model, apareixen els **models de intel·ligència artificial explicables** (de les seves sigles, XAI). Aquests models tenen com a principal objectiu millorar l'enteniment dels resultats obtinguts, fent que aquesta sigui capaç d'explicar com ha arribat a la sortida retornada. En aquest projecte concretament, es vol fer ús de un seguit de llibreries per a preparar un classificador binari que implementi un algorisme explicable i avaluar si les explicacions que genera l'algorisme són entenedores per als usuaris de peu.

Per posar un exemple, imaginem la situació de una intel·ligència artificial entrenada per al diagnòstic de càncer, la qual requereix com a entrada de dades una radiografia del pacient. A partir d'aquí, tindríem dues possibilitats: la primera és que el model anterior no sigui un model XAI. En cas d'introduir una radiografia com a entrada de dades, únicament rebríem una sortida numèrica, indicant si el diagnòstic que ha realitat la màquina és positiu (s'ha detectat un càncer) o negatiu (no s'ha detectat cap càncer). Com a segona opció, si el model anterior és un model XAI, a part de rebre aquesta sortida binaria indicant el resultat final, també rebríem a més una explicació o procés amb l'objectiu d'aconseguir un diagnòstic més clar, a part d'entendre bé quins passos o quin criteri ha aplicat el model per determinar el resultat final.

Hi ha un gran seguit d'algorismes que són capaços de generar diferents tipus d'explicacions, però aquest projecte es centrarà en un tipus d'explicació molt concret: els contrafactuals. Aquest tipus d'explicació consisteix en determinar els canvis que haurien de produir-se per a que una situació o esdeveniment que ha succeït canviés.

Durant el desenvolupament del projecte, es fa referència a certs termes que es mostren en català. La traducció del terme original en anglès ha estat extret del llibre *Guia sobre l'explicabilitat en la Intel·ligència Artificial* (Aussó et al., 2022).

1.1. Motivació

La principal motivació per a realitzar aquest projecte ha estat la gran utilitat que personalment trobo en aquests models. La intel·ligència artificial ha demostrat especialment en els darrers anys que és una tecnologia amb un gran rendiment, però sobretot molt extrapolable, ja que amb l'entrenament corresponent, pot ser útil en una gran quantitat de camps. Però des del punt de vista de la comprensió, és una tecnologia molt complexa d'entendre, ja que la majoria de models resulten ser una "caixa negra", per la qual cosa entendre completament que està passant a dins és quasi impossible, inclús per als desenvolupadors de software i científics responsables de la creació del model.

Gracies als models XAI, a part de rebre un resultat, també rebem una explicació del resultat obtingut. Amb aquest canvi, podem aconseguir que sigui molt més fàcil d'entendre de cara al usuari, el qual en la gran majoria dels casos pot no interpretar correctament un resultat sense l'acompanyament d'una explicació.

1.2. Objectius

Aquest projecte compta amb dos apartats. L'objectiu principal que es planteja per a l'apartat teòric és donar una pinzellada sobre què són els models d'intel·ligència artificial explicables, detallant els diferents aspectes i quins són els interessos darrere el seu estudi. També es donarà una vista als contrafactuals, el tipus d'explicacions en el que ens hem centrat.

Per a la part pràctica del projecte, l'objectiu serà realitzar una prova d'avaluació amb usuaris per avaluar la qualitat de les explicacions contrafactuals que retornarà un classificador binari que prèviament es prepararà, a més de comprovar si els usuaris es senten més satisfets rebent la informació en format taula o bé es senten més satisfets rebent la mateixa explicació però amb format text.

Per a poder assolir aquest objectiu, haurem de complir un requisit inicial: preparar el classificador. Aquest classificador binari rebrà com a entrada les dades de un client que sol·licita un crèdit i, primer de tot, generarà una predicció del risc de oferir aquest crèdit en qüestió. Donat que és un classificador binari, únicament podrà retornar dos valors: denegat i concedit. A més, per a aquells casos en que la predicció resulti en denegada, el classificador generarà un seguit d'explicacions contrafactuals, que indicaran al usuari quines dades de la sol·licitud han de canviar per a que el crèdit canviï el seu estat de denegat a concedit.

1.3. Estructura del projecte

Com hem explicat amb anterioritat, l'estructura del projecte tindrà dos pilars fonamentals: primer de tot, es realitzarà un apartat teòric, en el que es s'exposarà una visió general respecte què són els models d'intel·ligència artificial explicables, quins interessos hi ha al darrere de la implementació d'aquests sistemes, a més de un apartat centrat en les explicacions contrafactuals.

El segon pilar del projecte és un apartat pràctic, en el que es prepararà un classificador binari capaç de donar una predicció respecte el risc de concedir un crèdit a un client o no. A més, en cas de classificar-se com a denegat, oferirà unes quantes explicacions contrafactuals que mostrarà noves instàncies de la sol·licitud, amb les modificacions mínimes per a que, en el context d'aquest projecte, canviï l'estat de la sol·licitud de denegat a concedit. Per preparar aquest classificador, es farà ús de llibreries específiques de Machine Learning per entrenar un model amb un conjunt de dades de sol·licituds de crèdit, que realitzarà la classificació de les sol·licituds segons el seu risc. Posteriorment, mitjançant algorismes explicatius, es generarà un seguit d'explicacions contrafactuals d'aquelles sol·licituds que el model ha classificat com a risc alt.

Un cop es compleixi el requisit anterior, començarà l'objectiu principal del desenvolupament pràctic. Es proposarà un disseny experimental amb usuaris on es cercarà avaluar quin format millora la comprensió i satisfacció de l'usuari. Els dos formats que es plantegen són el format tabular, on es mostrarà una taula amb les explicacions contrafactuals o bé el format text, on s'exposarà la informació de la taula en format text.

2. Planificació del projecte

Tenint en compte que tenim un temps limitat per a realitzar el projecte, serà molt important organitzar el temps que es destinarà a cada un dels apartats del projecte. Primer de tot, inclourem una fase de documentació, en la que es destinarà la totalitat del temps a la lectura d'articles. Malgrat incloure aquesta fase especifica de lectura, en totes les fases del projecte es destinarà el temps necessari per llegir més articles o bé documentació en el cas de la preparació del classificador.

Per a la fase del desenvolupament de l'apartat pràctic, es destinarà una gran part del temps a la preparació del model, acompanyat de la lectura de la documentació de les llibreries. També es seguirà destintant temps a la lectura d'articles per a preparar el disseny experimental.

Finalment, es realitzarà de forma setmanal una reunió amb la tutora, moment en el que es plantejaran tots els dubtes que hagin pogut sorgir durant la setmana, ja sigui a nivell teòric o a nivell pràctic, o bé la definició de les tasques per dur a terme durant la setmana següent. A més, per a cada reunió es realitza un registre escrit de totes les tasques a realitzar durant la setmana pròxima, dubtes resolts i altre informació d'interès. Finalment, per a portar un registre de tot el procés, es compta amb un document separat per subapartats en el que s'exposen tots els problemes sorgits durant la preparació del model, resums d'articles.

2.1. Pressupost del classificador

Un factor important de la planificació del projecte serien els costos de preparar el classificador plantejat per a aquest projecte. Per fer aquest càlcul de temps, es plantejarà l'aproximació de hores dedicades a preparar el model i es realitzarà una multiplicació per el cost per hora. Per

extreure aquest càlcul, s'ha extret el sou mitjà de un programador de una publicació de la plataforma *Kiwi Remoto*, una plataforma destinada a la cerca d'ofertes de treball. Aquesta publicació (Kiwi Remoto, 2024) calcula una mitja del sou de totes les ofertes de treball publicades a la web i mostra que el sou mitjà de un desenvolupador és de **27€ per hora**. Tenint en compte això, el cost del projecte es calcula de la següent forma:

Fases de la preparació	Hores aproximades destinades	Cost	Cost total
Fase d'estudi i lectura	85 hores	27€/hora	2295€
Fase de preparació del model	192 hores	27€/hora	5184€
Total	277 hores	-	7.479€

Taula 1: Anàlisi i càlcul dels costos del classificador. Font: elaboració pròpia

A més, també hem de tenir en compte els costos associats a l'entorn de treball. Per calcular això, s'ha decidit plantejar la situació d'haver llogat una oficina de *coworking*, que consisteix en llogar unes instal·lacions compartides. Per donar un càlcul aproximat, s'ha realitzat un anàlisi del cost mensual de llogar una oficina a la plataforma COWORKING SPAIN (CoworkingSpain, 2024) que ofereix un gran seguit d'opcions. En aquesta pàgina, el preu mínim per llogar una oficina a Barcelona és de **180€ mensuals.** Tenint en compte el temps que s'ha disposat per a fer aquest treball ha estat 4 mesos aproximadament, el cost total ascendeix a **720€**.

Concepte	**Preu**
Pressupost del classificador	8199€

Taula 2: Pressupost final del classificador. Font: elaboració pròpia

3. Apartat teòric

En aquest apartat del projecte, es donarà una visió general de què són els models d'intel·ligència artificial explicables, a l'hora que s'explicarà l'interès que hi ha al darrere de la investigació d'aquests models i exposant què és una explicació, centrant-nos especialment en un tipus molt concret d'explicació: les explicacions contrafactuals.

3.1. Què és un model d'intel·ligència artificial explicable?

Per poder fer-nos una idea de què és un model d'intel·ligència artificial explicable, primer hem d'entendre que definim com al tipus de model contrari a un model explicable. Un **model opac** (en: *Black box model*) consisteix en un model d'intel·ligència artificial el qual no mostra en cap moment quins processos està realitzant en el seu interior. Això vol dir que, en el moment que un usuari ha introduït unes dades en el sistema i rep una sortida, aquesta sortida no inclou cap justificació ni indicació de com s'ha arribat a aquest punt, i a més no és lo suficientment interpretable com per entendre els processos que realitza en el seu interior. La gran majoria de models, donada la seva naturalesa, es poden classificar en aquest grup.

Figura 1: Esquema de un model opac. Entren unes dades i es genera una sortida, sense saber quin ha estat el procés.. Font: elaboració pròpia

Sabent això, podem començar a definir què és un model d'intel·ligència artificial explicable. Aquest consisteix en un tipus especial de model que és capaç de generar una explicació o mostrar el procés realitzat a l'hora de generar una predicció. Segons com realitzen una explicació de la predicció realitzada, parlem de diferents classificacions de models explicables.

Primer de tot, tenim els **models transparents** (en: *transparent models*). Aquest tipus de models es consideren "transparents" ja que el seu funcionament és molt interpretable, o sigui, que la seva estructura és molt simple d'entendre, com per exemple arbres de decisions de mida petita o bé models lineals (Christoph Molnar, 2022).

Per a aquests models, el seu principal avantatge també es converteix en el seu desavantatge més gran: aquests models tenen una estructura molt bàsica i fàcil d'entendre, factor que els dona més interpretabilitat respecte altres models com les xarxes neuronals, però també implica que el seu rendiment sigui molt menor envers a altres tipus de models.

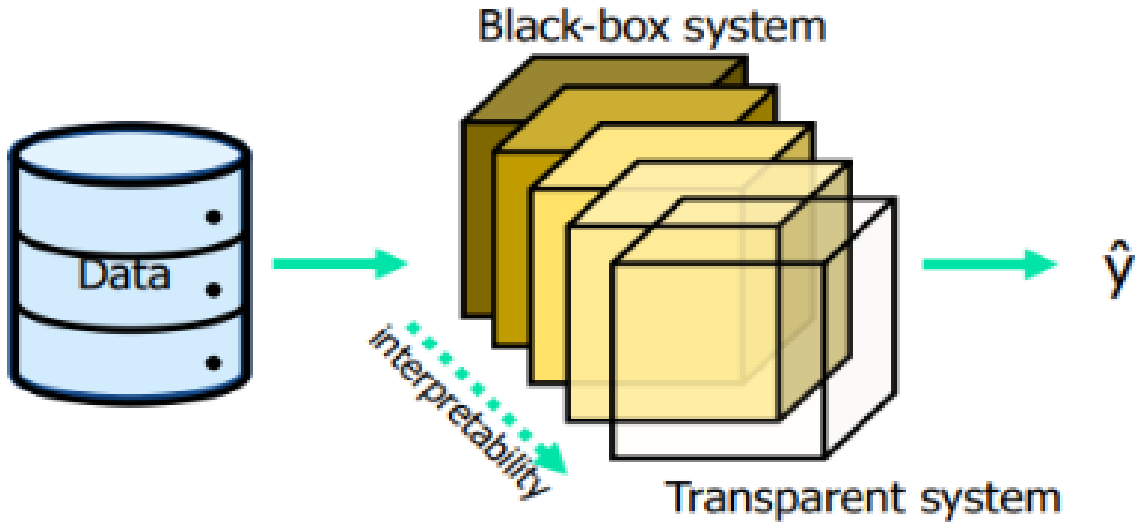

Figura 2: Esquema d'un model transparent. Com més transparent és, més interpretable és el seu funcionament
Font:(Tallaswapna, 2023)

Per a un altre banda, també podem trobar models interpretables mitjançant **explicabilitat post hoc** (en: *post hoc explanation methods*). Aquest sistema consta primer de tot d'un model opac convencional el qual és l'encarregat de generar una sortida mitjançant unes dades d'entrada que rep. Després, compta amb un sistema independent del model encarregat de generar l'explicació de com el model opac ha arribat al valor que ha retornat.

Dins d'aquest grup de tècniques explicables, podem trobar dos tipus d'algorismes:

- **Mètodes locals (en: local methods):** aquests mètodes consisteixen en donar una explicació de per què el model ha retornat una predicció per una instància concreta. Són molt útils per comprovar quines són les característiques del input que són més importants a la hora de donar una valoració.

- **Mètodes globals (en: global methods):** són aquells mètodes que permeten donar una explicació de quines son les característiques que tenen més impacte en el moment de que el model genera una sortida. Com indica Susanna Aussó (Aussó et al., 2022), aquestes tècniques descriuen el comportament a grans tres d'un model.

Christoph Molnar (Christoph Molnar, 2022) indica com a principal avantatge que aquest tipus d'algorismes donen molta llibertat als desenvolupadors, ja que l'explicabilitat post hoc és compatible amb qualsevol model, per la qual cosa el desenvolupador té la llibertat de poder escollir qualsevol model. A més, també destaca que aquestes tècniques poden ser aplicades per a models interpretables. Meike Nauta (Nauta et al., 2023) exposa que aplicar-li una tècnica explicativa a un model transparent com si es tractés de un model opac pot servir a més com a mètode d'avaluació, ja que, donat que el raonament d'un model transparent és conegut, l'explicació obtinguda per l'algorisme explicatiu pot ser comparada amb el raonament del model.

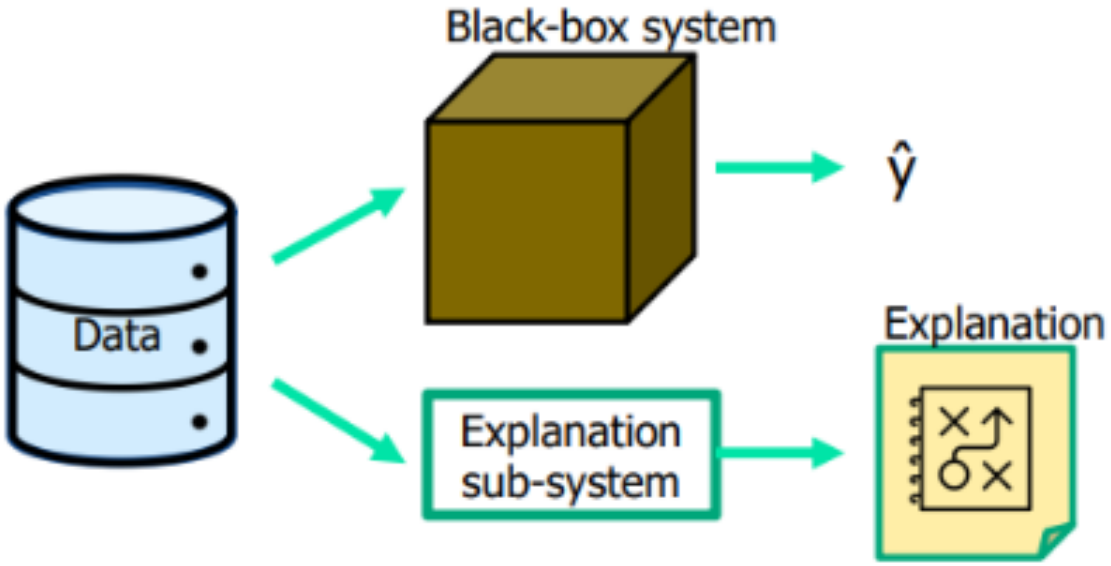

Figura 3: Esquema d'un model opac que utilitza una tècnica explicativa. Es pot observar com l'algorisme explicador es troba situat en un subsistema a part del model. Font: (Tallaswapna, 2023)

Tenint en compte aquesta llibertat a la hora d'escollir el model, podem determinar que gràcies a aquest tipus d'algorismes generadors d'explicacions podem implementar sistemes explicables que a la vegada no perdin rendiment degut a la senzillesa del model, per la qual cosa ens permetria implementar el millor dels models opacs (un gran rendiment) i dels models explicables (una explicació per entendre el resultat).

Aquests mètodes, malgrat semblar la millor opció, també compten amb un seguit de desavantatges. El primer desavantatge que podem trobar és el consum computacional, el qual és més elevat que el consum de un model interpretable. Això es pot veure de forma simple, ja que al utilitzar un mètode explicatiu, requerim del cost computacional de propi model, el qual pot tenir un consum més o menys elevat segons el tipus de model que usem, però també requerim del cost computacional del mètode que genera les explicacions, que novament el seu consum pot variar segons el mètode que escollim.

A més, hem de tenir en compte un factor important de les tècniques explicatives: l'algorisme que genera les explicacions és un component a banda del model opac, com podem veure a la **¡Error! No se encuentra el origen de la referencia.**. Això pot provocar que l'algorisme no sigui capaç d'entendre suficientment bé el model i, per tant, que el mètode generi explicacions que no coincideixen amb el raonament que ha realitzat el model i no acabin de ser precises.

	Models transparents	Models interpretables mitjançant explicabilitat post hoc
Avantatges	• Tenen una interpretabilitat molt alta degut a la seva estructuració simple. • Són fàcils d'implementar i requereixen menor temps d'entrenament.	• Ofereixen llibertat al desenvolupador a la hora d'escollir el model a utilitzar. • Al poder usar qualsevol model, si usem un model complex no perdem rendiment.
Desavantatges	• Rendiment baix degut a la seva estructura simple.	• Té un cost computacional més gran ja que requereix de un model i a més de un algorisme apart. • Donat que són dos sistemes per separat, el mètode explicatiu no sempre aconsegueix proporcionar una explicació precisa.

Taula 3: Taula d'avantatges i desavantatges de cada tipus de tècnica interpretable. Font: elaboració pròpia

3.2. Interès dels models d'intel·ligència artificial explicables

En els temps actuals s'han desenvolupat una gran quantitat de models d'intel·ligència artificial amb grans capacitats de resoldre problemes molt complexos. Però malgrat tenir models amb molt rendiment, es segueix fent recerca en aquest camp. Llavors, per què s'està fent tanta recerca en el camp de la intel·ligència artificial explicable?

- **Transparència**: actualment podem trobar aplicacions de la intel·ligència artificial en molts àmbits de la vida, i alguns d'aquests àmbits tenen una gran responsabilitat o bé uns grans conseqüències en cas que la predicció no sigui encertada. A més, en planteja un problema ètic bastant important. Qui té la responsabilitat de les decisions que pren un model en cas que aquestes siguin errònies? Andrea Berber (Berber & Srećković, 2023) indica que resulta molt complicat assignar responsabilitats ja que el model retorna una sortida sense que ningú tingui un control directe, a part que hi ha molts factors que poden portar aquesta decisió errònia i tots els factors no depenen d'una sola persona. Especialment en el camps que tenen més responsabilitat, com en la detecció de malalties, és molt important que el model sigui el més transparent possible a l'hora d'explicar els passos realitzats o en demostrar el criteri aplicat per a facilitar que la decisió presa per la màquina sigui adoptada per part de l'expert.

- **Augmentar la confiança**: ja sigui en el camp de la informàtica o qualsevol altre camp, si a l'hora de donar un resultat aquest no ve acompanyat d'una explicació del procés realitzat sempre es genera una desconfiança respecte el resultat. Per això, és interessant que el propi sistema sigui capaç de generar una explicació del resultat: en primer lloc, podem aconseguir augmentar la confiança en el públic objectiu, fent que sigui molt més fàcil introduir sistemes intel·ligents a la vida quotidiana. Per una altra banda, també pot

resultar molt útil per a l'equip de desenvolupament, ja que el rebre informació del procés que ha realitzat o bé una explicació, pot ajudar molt a detectar errors en el model o a proposar millores.

- **Millora dels algorismes:** a la hora de desenvolupar els algorismes i els models de intel·ligència artificial, és molt important per als desenvolupadors entendre la lògica que hi ha darrere les sortides que retorna el model per a poder corregir els possibles errors o bé desenvolupar millores. Gràcies a les explicacions del model, aquest treball d'entendre els resultats es facilita en gran mesura.

- **Ètica:** al llarg del temps les persones han pres decisions importants molt influenciades per certs prejudicis, ja siguin racials o de gènere per exemple. Com indica la entrada al bloc del Institut d'Enginyeria del Software (SEI) (Violet Turri, 2022), els sistemes intel·ligents de forma no intencional, acaben desenvolupant els mateixos prejudicis que han desenvolupat les persones, degut principalment a que les dades que s'utilitzen per entrenar els models reprodueixen aquests biaixos. Per corroborar aquesta afirmació, ens podem adreçar a l'exemple descrit en un estudi de la revista Forbes (Hale, 2021), el qual afirma que els biaixos produïts per les intel·ligències artificials causa que el 80% dels sol·licitants de crèdit de races negres acaben sent denegats malgrat compartir moltíssimes característiques amb altres perfils de gent blanca. Els models explicables, a nivell teòric, poden facilitar la correcció de biaixos donat a que amb les explicacions retornades, simplifica el detectar certs comportaments durant el seu desenvolupament, permetent als desenvolupadors corregir-los.

3.3. Definició d'explicació

Al llarg del projecte s'ha mencionat i es mencionarà més d'un cop la següent paraula: explicació. Però realment sabem respondre a què és una explicació? Si ens adrecem al Diccionario de la lengua española podem trobar les següents definicions (Real Academia Española, 2023) :

"Declaración o exposición de cualquier materia, doctrina o texto con palabras claras o ejemplos, para que se haga más perceptible"

"Satisfacción que se da a una persona o colectividad declarando que las palabras o actos que pueden tomar a ofensa carecieron de intención de agravio"

"Manifestación o revelación de la causa o motivo de algo"

Figura 4: Definicions de explicació exposats al Diccionario de la lengua española.

Com podem veure, de una sola font d'informació hem pogut extreure tres possibles definicions d'explicació, i cap d'elles resulta errònia, però hem de tenir en compte que no totes s'adapten al context en el que estem treballant.

Per exemple, posem que tenim un model explicable que rebent com a entrades un seguit de radiografies de una persona és capaç de donar un percentatge de probabilitat de que tumor sigui maligne. En aquesta situació, l'explicació vol fer entendre al pacient els motius per els quals el model ha donat el percentatge de probabilitat corresponent.

Si analitzem les diferents definicions exposades anteriorment, podem veure que:

- La primera definició no s'acaba d'ajustar del tot al context de la intel·ligència artificial explicable, ja que l'objectiu del model no és fer que un temari o matèria sigui més perceptible per el usuari, sinó que es vol aconseguir que l'usuari entengui bé el resultat obtingut.
- La segona explicació no s'ajusta en res, ja que en cap moment el model vol justificar un ofensa que hagi realitzat el sistema cap a un col·lectiu determinat, encara que com s'ha indicat en la secció anterior, el sistema pugui produir certs prejudicis cap a cert sector dels usuaris.
- La tercera definició podria ser la que millor s'ajusta. Com indica la definició, una explicació és el motiu d'alguna cosa i l'objectiu del model és justificar el motiu per el qual ha donat una predicció a un problema

Veient que l'explicació és un dels termes més importants darrere els models explicables, definirem en aquest projecte el terme *explicació* de la següent forma, tenint en compte que en aquest projecte ens centrarem en les explicacions als usuaris no especialitzats:

> "Intenció de justificar un criteri aplicat a la hora de resoldre un problema o bé els passos seguits a la fi de poder aclarir els dubtes o inquietuds produïdes a un usuari sense coneixements previs respecte al camp tractat"

Figura 5: Definició pròpia orientada al context que tractem

Un altre terme que apareix en el camp de la intel·ligència artificial és la interpretabilitat de un model. Aquest terme molts cops es confon amb el terme explicació, i malgrat poden tenir certa similitud, no té el mateix significat.

Si cerquem novament al Diccionario de la lengua española, podrem veure que els significats són els següents (Real Academia española, 2023):

> "Explicar o declarar el sentido de algo, y principalmente el de un texto"
> "Traducir algo de una lengua a otra, sobre todo cuando se hace oralmente"
> "Explicar acciones, dichos o sucesos que pueden ser entendidos de diferentes modos"
> "Concebir, ordenar o expresar de un modo personal la realidad"

Figura 6: definicions de interpretar exposades a la Real Academia Española

De la mateixa forma que a la definició d'explicació, el context en el que treballem també fa variar el significat d'interpretació. Seguint amb la situació exposada en el punt anterior, definirem la interpretabilitat de la següent forma:

"Capacitat d'entendre o traduir quin és el funcionament de un sistema tenint en compte els resultats obtinguts"

Figura 7: definició d'interpretar per a aquest projecte

3.3.1. Motius de una persona per demanar una explicació

Els motius per el quals una persona pot demanar una explicació d'un fenomen o resultat poden ser molt variats. Tim Miller menciona a l'article *Explanation in artificial intelligence: Insights from the social sciences* (Miller, 2018) que "la curiositat és una dels criteris primaris que usa les persones, però altres motius pragmàtics també inclouen examinació – per exemple, quan un professor pregunta als seus alumnes per una explicació en un examen amb l'objectiu de posar-los a prova; i explicacions científiques – preguntant-se per què observem un fenomen ambiental particular.

La base dels motius per els quals una persona pot requerir una explicació és **l'enteniment**: una persona vol millorar la comprensió de, per exemple, un esdeveniment que acaba de succeir del qual no té coneixements previs o bé no disposa dels coneixements suficients com per a entendre-ho per si mateixa. D'aquesta base podem extreure posteriorment altres motius, com per exemple **l'aprenentatge** - una persona té pocs o cap coneixement sobre un temari o sector i vol augmentar-los per diferents motiu; o bé la **resolució de problemes** – una persona es troba davant un problema que ha de resoldre, però degut a la manca dels coneixements necessaris no sap com plantejar una possible solució.

3.3.2. Complexitat de una explicació

També hem de tenir en compte, però, un altre factor molt determinant de una explicació: la seva complexitat. Una explicació ha de tenir com a objectiu esclarir alguna cosa, però si el contingut d'aquesta explicació no s'ajusta al nivell de coneixement que té el públic objectiu, l'efecte que aconseguim és al contrari que volem aconseguir amb una explicació, ja que amb explicacions molt tècniques podem generar més dubtes que els inicialment plantejats al usuari, augmentant la inseguretat al respecte.

Seguint l'exemple anteriorment exposat, l'explicació que rebrà un pacient per a entendre el percentatge de risc de que el tumor detectat sigui maligne no pot tenir el mateix nivell de complexitat que l'explicació que rebrà un metge titulat, ja que és molt probable que els coneixements amb el que compte el metge siguin molt superiors als coneixements del pacient.

3.4. Tipus d'explicacions

Un punt molt important de les explicacions és la gran quantitat de tipus d'explicacions. Com indica Lombrozo (Lombrozo, 2012), no totes les explicacions tindran necessàriament la mateixa estructura o funció. Per exemple tenim les explicacions causals que cerquen respondre als motius reals de un esdeveniment, explicacions descriptives que cerquen explicar què és un element, entre altres. En aquest projecte ens centrarem en un tipus molt concret d'explicacions, les explicacions contrafactuals.

3.5. Explicacions contrafactuals

Per entendre bé què són les explicacions contrafactuals, haurem d'entendre bé què és una predicció. Com es defineix en el llibre *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Christoph Molnar, 2022), el valor de una predicció ve donada per un seguit de característiques de una instància.

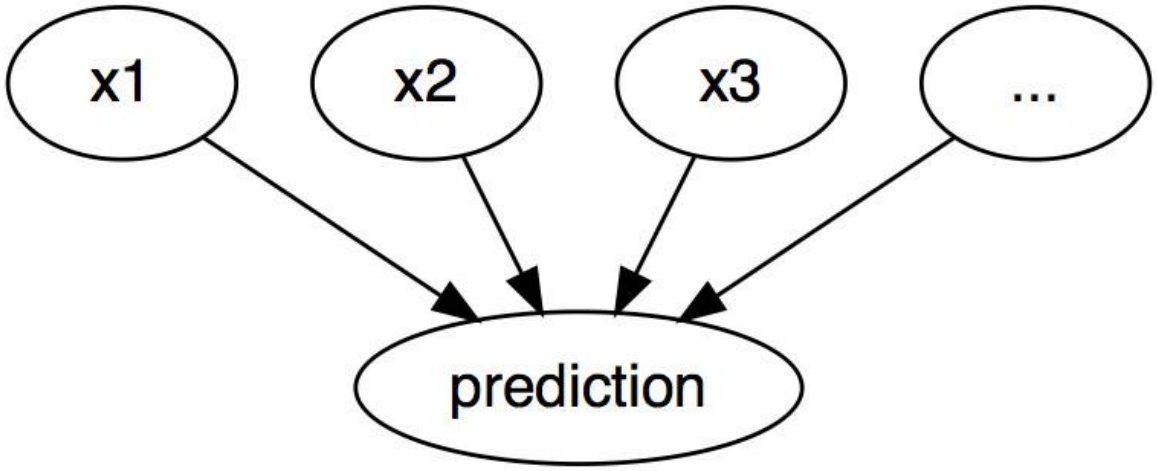

Figura 8: representació en forma de graf de la relació entre predicció i input (x1, x2, ...). Font: (Christoph Molnar, 2022)

Tenint en compte això, una **explicació contrafactual** consisteix en el canvi mínim que s'ha de produir en les característiques de X per a que el resultat de la predicció canviï de forma dràstica, on la predicció pot ser un esdeveniment que ha succeït degut a certs factors o bé una decisió.

Aquest tipus d'explicacions resulten ser molt útils en el camp de la intel·ligència artificial explicable. Hem de tenir en compte que molts dels models tenen com a objectiu classificar instàncies de dades en classes. Gràcies als contrafactuals, podem aconseguir veure quines són aquelles característiques de la instància que haurien de canviar per a que el model classifiqués la instància en una altra classe.

Per posar un exemple, imaginem la situació de una persona que sol·licita un crèdit a un banc. Aquest banc demana unes dades personals de la persona per a poder prendre una decisió sobre si el risc de oferir el crèdit a aquesta persona és assumible, o bé és massa alt. Les dades que un banc demana a un sol·licitant de un crèdit solen consistir, entre altres, en:

- Edat
- Lloc de treball
- Import de crèdit
- Finalitat del crèdit
- Termini de devolució

Un cop la persona ha omplert les dades, el banc comença a analitzar-les. Passat un cert temps, la persona rep una notificació del banc conforme el crèdit ha estat denegat. Llavors, la persona decideix visitar el banc per rebre una explicació del motiu per el qual s'ha denegat el crèdit. Llavors, l'assessor de crèdits explicarà quins són els motius, o dit d'un altre manera, mencionarà quines característiques de la persona han de canviar per a que el crèdit s'aprovi.

Si intentem plantejar aquesta situació real seguint l'explicació anterior, veurem el següent:

1. La predicció en aquest cas concret és la classificació que ha realitzat el model, o sigui, si el crèdit ha estat denegat o bé concedit per el banc.
2. Les característiques o inputs equivalen a les dades que ha recollit el banc en el moment que s'ha sol·licitat el crèdit.
3. Finalment, el banc podria facilitar al client què hauria de canviar per a que la predicció canviés de denegat a concedit. Aquest tipus d'explicacions són el que anomenem *contrafactual*.

Hi ha certes maneres d'avaluar si un contrafactual és bo. Christoph Molnar (Christoph Molnar, 2022) considera que un bon contrafactual ha de seguir un llistat de característiques.

La primera característica que ha de tenir un bon contrafactual és que ha de ser el més similar possible a la instància original, o sigui, en el moment que es genera el contrafactual, aquest ha de canviar **el mínim possible** per a que pugui ser considerat bo. Per calcular la diferència entre dos instàncies, podem fer ús de la distància de Manhattan o la distancia Gower (Christoph Molnar, 2022).

Un altre característica que ha de tenir un bon contrafactual és que **els valors que s'indiquin que han de variar s'han de poder modificar**. Per entendre aquest punt, exposarem un altre exemple exposat per Christoph Molnar (Christoph Molnar, 2022): imaginem que disposem de un pis que volem llogar. Llavors comptem amb un sistema que ens indica el preu del lloguer segons les característiques del pis. Introduïm les dades al sistema i ens diu que el preu del lloguer del pis és de 900€. Llavors preguntem al sistema que hauria de canviar per poder llogar-lo per 1000€. En aquesta situació haurem de revisar que el contrafactual no pugui generar valors sense sentit: per exemple, un contrafactual no ha de retornar un valor de superfície del pis major que la que hi ha, ja que els metres quadrats del pis són fixes i no es poden modificar.

A més, a fi de poder garantir una bona explicació contrafactual, és recomanable que es generin **més de una explicació contrafactual**, per a que el model tingui més d'una possible via per

generar una sortida diferent. Clara Bove (Bove et al., 2023) mostra en el seu estudi que l'ús de més de una explicació contrafactual ajuda als usuaris a entendre millors les explicacions.

Hem de tenir en compte, però, que els contrafactuals poden patir un efecte conegut com **efecte Rashomon** (Anderson, 2016). En termes generals, consisteix en un efecte en que dos persones justifiquen una mateixa situació de forma diferent sense que cap de les dues sigui incorrecte, a causa de la subjectivitat de cada un dels individus a la hora d'entendre la situació. Això mateix pot passar amb els contrafactuals. Com menciona Christoph Molnar (Christoph Molnar, 2022), "cada contrafactual diu una historia diferent de com s'ha arribat a una sortida determinada. Un contrafactual pot indicar que s'ha de canviar una característica A i un altre pot indicar que la característica A s'ha de deixar igual i només s'ha de canviar una característica B".

4. Apartat pràctic

En aquest apartat de la memòria s'exposarà quin han estat els passos per a realitzar la preparació del model classificador. Explicarem quin ha estat el tractament que s'ha realitzat de les dades, com s'ha configurat i entrenat el model i posteriorment s'exposarà el mètode que s'utilitzarà per avaluar quin format de contrafactuals resulta més comprensiu per els usuaris: format tabular o bé format textual.

4.1. Conjunt de dades

Les dades utilitzades en aquest projecte consisteixen en un data set format per mil entrades de persones sol·licitants de un crèdit bancari. Aquestes dades han estat extretes de Kaggle (Specified, 2017). Cal dir, però, que aquest conjunt de dades és una reducció de característiques de un altre conjunt de dades, present al Repositori de Machine Learning UC Irvine (Hans Hofmann, 1994).

4.1.1. Dades originals

El data set original compte amb un total de 11 columnes o camps. Primer de tot tenim el camp de **identificador**. Aquest cap consisteix en un identificador únic que ens permet identificar els usuaris dins del data set sense la necessitat d'altres dades delicades com noms o documents d'identitat.

Després tenim el camp **Age,** consisteix en la edat de la persona que sol·licita el crèdit, representat mitjançant un valor enter. La distribució de les edats en el data set són el següent:

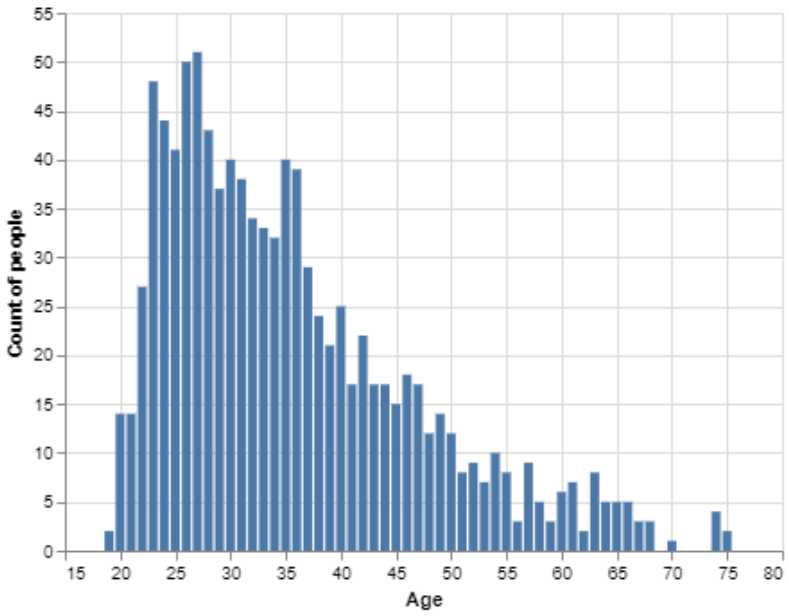

Figura 9: Distribució de l'edat en el data set. Font: elaboració pròpia

En el gràfic següent, podem veure la distribució de les edats dels clients que han sol·licitat el crèdit. Observem com la franja d'edat amb més clients és la franja que engloba els 20 anys fins els 35 anys, que conté un 53% de la mostra.

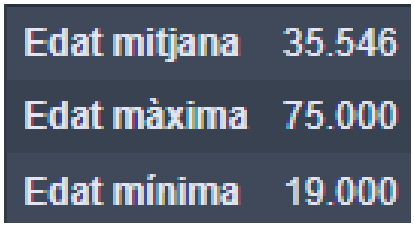

Figura 10: Edat mitjana, màxima i mínima del conjunt de clients. Font: elaboració pròpia

A la taula anterior es mostra quins són els resultats d'extreure un valor mitja, mínim i màxim de les edats dels clients. Com podem observar, els clients més joves que han sol·licitat un crèdit tenen 19 anys, mentre que l'edat més gran representada en aquest conjunt és de 75 anys, amb una mitjana d'edat d'exactament 35,546 anys.

La següent columna de la qual disposem és **Sex**, la qual indica quin sexe té la persona. Per representar-ho, tenim un valor categòric que pot prendre dos valors: male per al sexe masculí, o female per al sexe femení.

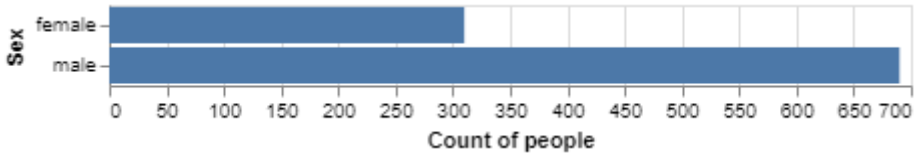

Figura 11: Distribució dels usuaris segons sexe. Font: elaboració pròpia

Podem veure en el gràfic anterior que el número de persones de sexe masculí volta cap a les 700 persones, lleugerament superior al doble de persones de sexe femení.
També comptem amb la columna **Housing**, la qual ens indica de quin tipus d'habitatge disposa el client. Aquest camp és un valor categòric que pot tenir tres possibles valors:

- Free: el client no disposa de cap habitatge
- Rent: el client disposa de un habitatge llogat.
- Own: el client compte amb un habitatge comprat.

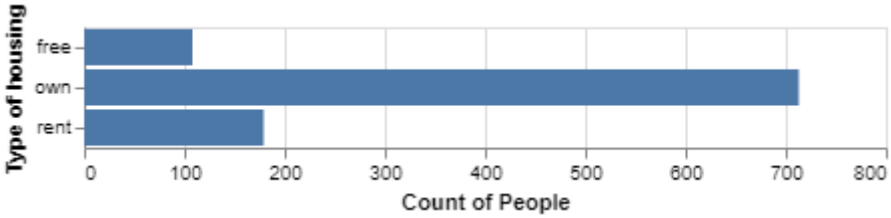

Figura 12: : Distribució dels clients segons el tipus d'habitatge. Font: elaboració pròpia

Segons el gràfic anterior, podem veure que del total de 1000 clients, aproximadament un 70% compten amb un habitatge comprat, mentre que aproximadament uns 200 clients es troben

vivint en un habitatge llogat. Els 100 clients restants no disposen de cap habitatge, ni llogat ni comprat.

La següent columna present a la taula és la columna **Saving accounts**. Aquest camp indica, per a aquells usuaris que compten amb comptes d'inversió, en quita categoria es classifica aquesta compte segons el seu valor. En el data set original (Hans Hofmann, 1994), classifica les comptes segons els valors mostrats a continuació, amb la diferencia que aquest valor numèric ve representat amb una moneda DM, la qual es substituirà a euros per facilitat l'enteniment. Els valors que podem trobar en aquesta columna són els següents:

- Little: compte amb un valor menor a 100€.
- Moderate: compte amb un valor entre 100 i 500€.
- Quite rich: compte amb un valor entre 500 i 1000€.
- Rich: compte amb un valor superior a 1000€.

La distribució de les dades és la següent:

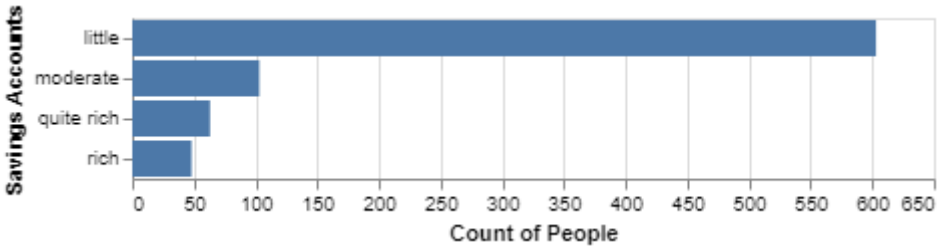

Figura 13: Distribució dels clients segons la compte d'estalvi. Font: elaboració pròpia

El gràfic mostra que la gran majoria dels usuaris compten amb un compte d'estalvi de nivell baix (aproximadament unes 600 persones de 1000, un 60%).

La següent columna del data set és **Checking account**, que indica quin nivell adquisitiu té el compte corrent del client, en cas que en tingui una. Aquest valor categòric pot tenir els següents valors:

- Little: el compte corrent té un nivell adquisitiu baix.
- Moderate: el compte corrent té un nivell adquisitiu mitjà.
- Rich: el compte corrent compte amb un nivell adquisitiu alt.

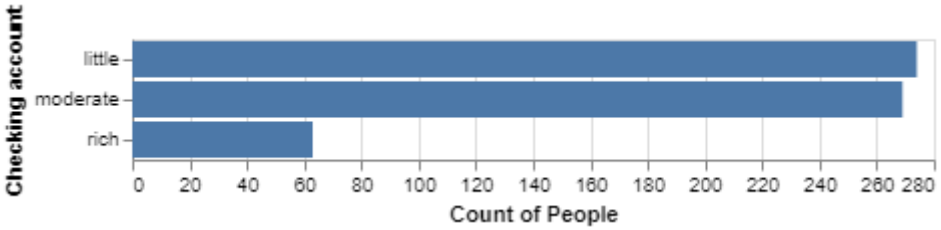

Figura 14: Distribució dels clients segons el nivell adquisitiu del compte corrent. Font: elaboració pròpia

	Checking account	Numero persones	Percentatge
0	little	274	0.274
1	moderate	269	0.269
2	rich	63	0.063

Figura 15: Percentatge de persones per a cada tipus de compte. Font: elaboració pròpia

Com podem veure a les il·lustracions anteriors, la gran majoria dels clients compten amb un compte corrent de nivell baix o moderat, mentre que 63 persones compten amb un compte de nivell alt. Si ens fixem, però, podem veure que hi ha un restant 394 clients que no compten amb cap compte corrent, o sigui, en el data set es mostra com Nan.

La següent columna és la columna **Credit amount**. Aquesta columna conté un valor enter que indica la quantitat de crèdit que ha sol·licitat el client.

Valor del credit mitjà	Valor de crèdit màxim	Valor de crèdit mínim
3271.258	18424	250

Figura 16: Valor mitjà, mínim i màxim de la columna crèdit. Font: elaboració pròpia

Com indica la imatge anterior, podem veure que el crèdit més petit sol·licitat ha estat de 250, el més gran té un valor de 18424 i el crèdit mitjà té un valor de 3271.258.

Després, també tenim la columna **Duration**, la qual ens indica amb un valor enter el número de mesos en el que el client vol retornar el crèdit sol·licitat.

Temps mitjà	Temps mínim	Temps màxim
20.903	4	72

Figura 17: Temps mitjà de retorn, temps mínim i màxim. Font: elaboració pròpia

Si mirem la taula, veiem que el temps mitjà de duració és de 20.903 mesos, amb un valor mínim de 4 mesos i un valor màxim de 72 mesos.

La següent columna és **Purpose**, la qual ens indica quin és el motiu per el qual es sol·licita el crèdit. Aquesta columna categòrica pot tenir qualsevol dels valors següents:

- Radio/TV: indica que el motiu és l'adquisició de una televisió o ràdio.
- Education: indica que el motiu és una inversió en educació (curs, universitat,...)
- Furniture/equipment: indica que el motiu és adquirir mobles o altre equipament per a la llar.
- Car: indica que el motiu és adquirir un vehicle.

- Business: indica que el motiu és per realitzar una inversió en un negoci.
- Domestic appliances: indica que el motiu és adquirir maquinaria per a la casa ja sigui per cuinar o bé per salut.
- Repairs: indica que el motiu és realitzar reparacions
- Vacation/others: agrupació de motius que inclou pagar vacances o altres motius.

La distribució dels clients segons el motiu de la sol·licitud és la següent:

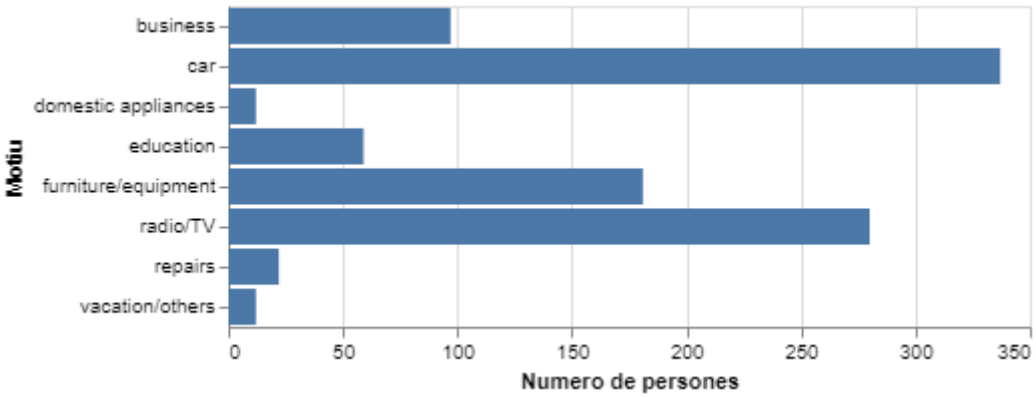

Figura 18: Distribució dels clients segons el motiu de la sol·licitud. Font: elaboració pròpia

Purpose	Numero persones	Percentatge
business	97	0.097
car	337	0.337
domestic appliances	12	0.012
education	59	0.059
furniture/equipment	181	0.181
radio/TV	280	0.280
repairs	22	0.022
vacation/others	12	0.012

Figura 19: Percentatge de clients per cada motiu. Font: elaboració pròpia

Com podem veure a les imatges anterior, el motiu principal per sol·licitar un crèdit és adquirir un vehicle (amb un 34% dels clients aproximadament), seguit per adquirir una televisió o una ràdio (un 28% dels clients) i adquisició de mobles i equipament (un 18% dels clients).

La ultima columna que troben en el data set és la columna **Risk**. Aquesta columna categòrica és el nostre target, o sigui la variable la qual volem predir, i indica si el risc de oferir un crèdit al client és positiu o negatiu. Aquesta columna únicament pot prendre dos valors:

- Good: indica un risc positiu.
- Bad: indica un risc negatiu.

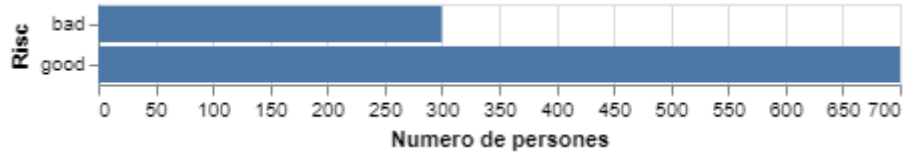

Figura 20: Distribució dels clients segons el risc de la sol·licitud. Font: elaboració pròpia

Risk	Numero persones	Percentatge
bad	300	0.3
good	700	0.7

Figura 21: Percentatge de la distribució dels clients. Font: elaboració pròpia

Com podem veure a les imatges anteriors, tenim un total de 300 sol·licituds que han estat rebutjades i un total de 700 sol·licituds acceptades (un 30% i un 70% respectivament).

4.2. La llibreria AIX360 i les seves possibilitats

Com s'ha exposat en el punt anterior, per al desenvolupament s'ha utilitzat AIX360, una llibreria de codi obert que permet implementar millores d'interpretabilitat i generació d'explicacions per a conjunt de dades i models de Machine Learning.

La llibreria en qüestió va estat implementada per l'empresa informàtica IBM Research, però donada a la fundació *Linux Foundation AI &* Data (The Linux Foundation, 2024) i es pot descarregar des de un repositori públic a GitHub (IBM, 2023). Actualment compte amb una comunitat activa, ja que si es revisa el repositori, es troben propostes de millora de la implementació dels algorismes, a l'hora que també es disposa d'un fòrum de incidències on es presenten diferents problemes trobats per els usuaris.

En el propi repositori, s'enllaça la pagina oficial de la llibreria (IBM research trusted AI, n.d.) on es mostra molta informació de la llibreria, com vídeos introductoris, un article que explica com es va dissenyar la llibreria i a més compte amb un seguit d'exemples en format notebook on s'exposen un seguit d'exemples de implementacions amb diferents mètodes inclosos a la llibreria.

A l'hora, també compte amb una API (IBM, 2019) on es troba tota la documentació necessària per a poder fer ús dels mètodes de la llibreria per generar explicacions, així com també funcionalitats per al càlcul de mètriques i finalment instruccions de com utilitzar els data sets preparats en el interior de la llibreria.

Com ens indica la documentació de la llibreria, disposem de un seguit d'algorismes diferents per generar explicacions, tan mètodes locals com mètodes globals. A més, també compta amb un seguit de funcions per a extreure mètriques per comprovar si l'explicador està interpretant el resultat del model de forma correcte. Els dos algorismes per extreure mètriques són els següents:

- **Faithfulness metric**: aquesta mètrica s'encarrega de fer una avaluació entre la correlació entre la importància que l'algorisme ha donat a cadascun dels atributs i l'efecte de cada atribut en el rendiment del model. O sigui, revisa que l'algorisme hagi assignat correctament el pes de cada atribut en la predicció del model (Alvarez-Melis & Jaakkola, 2018).

- **Monotonicity mètric**: aquesta mètrica s'encarrega de calcular l'efecte que tindrà cada un dels atributs. Per fer això, va calcular el rendiment del model primer amb un sol atribut, per posteriorment afegir la resta de un en un. Cada un dels atributs s'afegeix per ordre de importància (de menor importància a major) (Luss et al., 2019).

4.3. Eines utilitzades per a la preparació del classificador

Un cop escollit un llenguatge de molt alt nivell com és Python, s'havia d'escollir la eina que s'utilitzaria per a la preparació del classificador. Era molt important que la eina permetés treballar de forma òptima i no generés cap problema a la hora d'utilitzar les llibreries pertinents. Tenint en compte els següents punts, es van plantejar tres possibles opcions.

La primera opció va ser utilitzar **Google Colab**, una eina pròpia de Google la qual permet editar fitxers de tipus notebook sense necessitat d'instal·lar cap aplicació. Aquest programa permet editar fitxers que es trobin emmagatzemades a Drive, per la qual cosa es va proposar preparar una carpeta que contingués en el seu interior el fitxer notebook i els fitxers amb les dades necessàries.

Llavors ens vam topar amb un problema bastant complicat de resoldre: l'aplicació no permet guardar un entorn ni instal·lar llibreries amb la consola sense pagar una subscripció. Per resoldre-ho, es va intentar preparar una cel·la dins del mateix notebook la qual realitzés la instal·lació de totes les llibreries en el moment d'executar-la, però ens vam adonar que a la hora de instal·lar certes llibreries, com per exemple SKlearn, rebíem un error a la hora de instal·lar certes versions ja que aquestes no es trobaven disponibles. Tenint en compte que tan la llibreria Carla com AIX360 requereixen de una versió especifica de SKlearn que no es trobava disponible, es va descartar utilitzar Google Colab.

La segona eina que es va proposar fou **Jupyter-Hub**, un projecte de codi obert que permet crear entorns i fitxers accessibles des de més de un ordinador alhora. La idea era bastant prometedora, però no es va poder dur a terme ja que era necessari disposar de un servidor Linux amb nodejs i npm corrent en ell, junt amb un domini i un certificat SSL. Malgrat revisar-se diferents opcions de hosting per poder emmagatzemar aquesta instancia de Jupyter-Hub, no es va trobar cap opció raonable, per la qual cosa la eina es va descartar.

La tercera i ultima opció que es va rumiar va ser publicar el notebook en un repositori de **GitHub**, el qual ens permetria tenir accés als documents, però sense poder visualitzar-ho a temps real com s'havia pensat en un inici. Com a part positiva, donat que el fitxer es troba en un repositori, es compta amb la llibertat d'utilitzar l'editor de codi que es desitgi, de disposar de tots els fitxers

necessaris, ja siguin fitxers de codis o fitxers de dades, en una mateixa ubicació, a més de seguir mantenint la possibilitat de treballar des de diferents ordinadors.

Tenint en compte això, per al desenvolupament del projecte pràctic, s'ha decidit utilitzar com a eines principals Anaconda Navigator, una aplicació que ens permet gestionar de forma simple els entorns de desenvolupament, la instal·lació de paquets i el llançament d'aplicacions. Concretament, l'aplicació amb la que s'ha preparat el classificador binari ha estat Jupyter Notebook, projecte de codi obert molt útil per a la creació i edició de fitxers notebook.

Un cop escollit també l'editor de codi que utilitzarem, la següent iteració és escollir la llibreria que s'utilitzaria per generar les explicacions. En un inici, es va decidir utilitzar la llibreria Carla (Pawelczyk et al., 2021a), una llibreria de codi obert que ofereix un gran seguit d'algorismes per generar contrafactuals. Va ser desenvolupada per Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter i Gjergji Kasneci.

Un dels principals motius per decantar-se per la llibreria Carla és la gran quantitat de configuracions possibles: primer de tot, ens permet escollir entre tres models inclosos a la llibreria – xarxa neuronal, model lineal o bé un arbre de decisions – o bé, mitjançant una classe abstracta, podem implementar de 0 el nostre propi model. També disposem de un llistat bastant ampli de mètodes per extreure contrafactuals, podent escollir aquell que millor s'adapta al problema a resoldre.

Aquest llibreria es va trobar en ús durant aproximadament uns 3 mesos, i malgrat les facilitats que semblava aportar, finalment es va decidir descartar el seu ús, degut principalment als motius que s'exposen a continuació.

El primer problema no bloquejant que va aparèixer va ser el model que utilitzaríem. Com s'ha mostrat anteriorment, la llibreria compte amb un gran seguit d'opcions de configuració. Llavors, per comprovar quin dels models inclosos s'ajustava millor al nostre problema, es va decidir generar un model amb cada tipus de backend per comprovar quins valors de precisió obteníem. Primer de tot, es va preparar una xarxa neuronal, però malgrat realitzar proves amb els dos backend compatibles (donat que SKlearn i XGBoost no són compatibles), els resultats obtinguts per les dues implementacions eren equivalents, amb una precisió (en: *accuracy*) de 0'668.

La següent prova que es va realitzar va ser implementar un model lineal, que només és compatible amb els backends tensorflow i Pytorch. Els resultats obtinguts en el cas de tensorflow van empitjorar dràsticament, ja que la precisió que es va obtenir amb el mateix conjunt de dades era equivalent a 0.476, una diferència de 0.212, el qual no és mínimament acceptable en cap situació. En canvi, amb el backend Pytorch, el rendiment es mantenia força equivalent.

La ultima prova va ser implementar un model de Random Forest. Aquest model no és compatible amb els backends Tensorflow i Pytorch, però si amb SKLearn i XGBoost. La precisió obtinguda va ser per SKlearn un 0.684 i per XGBoost un 0.656, un rendiment molt similar entre els diferents backends, a més de un rendiment molt similar a la resta de models.

Per solucionar aquest problema de rendiment es podia configurar els paràmetres d'entrenament i millorar lleugerament aquests valors, però el problema principal va sorgir a la hora de generar els contrafactuals. Es van realitzar proves amb alguns dels mètodes disponibles i tots retornaven errors d'execució que no ens permetien generar els contrafactuals. Malgrat tots els esforços per intentar esbrinar si el problema podia ser degut a la codificació de les dades que era erroni o bé que el model no estava ben definit, no va ser possible degut a la falta de documentació: la llibreria compte amb una web on trobem tota la documentació dels mètodes (Pawelczyk et al., 2021b) però aquesta documentació no es troba actualitzada. A més, els recursos a internet eren molt reduïts.

Degut a tots aquests problemes exposats, es va decidir cercar un altre llibreria. Es va decidir començar a utilitzar una llibreria suggerida per Jordi Vitrià, AIX360, la qual també ofereix una gran quantitat de mètodes per generar explicacions. Després del procés d'instal·lació i provar els resultats que obtenia, es va veure que la llibreria era molt senzilla d'utilitzar, a part de comptar amb molta més documentació disponible per consultar i ofereix un gran ventall de possibilitats de configuració. Per això, AIX360 es va imposar com la llibreria utilitzada per aquest projecte.

4.4. Tractament de les dades

Si analitzem les dades presents dins el fitxer de dades, podem veure que únicament dues columnes contenen valors nuls o nan.

```
1  data = pd.DataFrame(pd.read_csv("./archive/german_credit_data.csv")).drop("Id", axis=1)
2  data.isnull().sum().sort_values(ascending=False)
```

```
Checking account    394
Saving accounts     183
Age                   0
Sex                   0
Job                   0
Housing               0
Credit amount         0
Duration              0
Purpose               0
Risk                  0
dtype: int64
```

Figura 22: Distribució de valors nuls segons cada columna del data set. Font: elaboració pròpia

Si mirem la il·lustració anterior, podem veure que a les columnes *Checking account* i *Saving Accounts* comptem amb 394 i 183 valors nuls respectivament. Es va topar amb que, a l'hora de preparar les dades per el model, , tan la llibreria Carla com la llibreria AIX360 retornaven un error conforme no accepta valors nuls, motiu per el que es va decidir fer un tractament previ de les dades, extraient aquests valors no vàlids.

La primera opció que es va analitzar va ser la possibilitat d'eliminar aquelles entrades que continguessin algun dels seus atributs amb un valor nul, però al veure el data set resultant, es va descartar ràpidament, ja que la reducció del número de columnes era massa gran.

La segona opció va ser la possibilitat de substituir aquests valors nuls per un altre valor. Tenint en compte que el ventall de possibles valors de les dues columnes era el mateix, es va plantejar afegir un nou valor possible a aquest ventall**.** Aquest valor en el valor ***unknown***. D'aquesta forma, s'evita perdre instàncies de dades.

Un cop tenim totes les dades preparades, haurem de tractar aquelles dades del conjunt que siguin **dades categòriques:** aquelles dades que es representen amb un cert rang de valors determinats i que classifiquen les files de dades. Per exemple, una dada categòrica podria ser situació laboral, ja que classifica a, en aquest cas un client, en dues categories, segons si es troba treballant o bé es troba a l'atur.

El problema que ha sorgit amb aquest tipus de dades és que el model utilitzat en el projecte no accepta aquest tipus de dades, ja que aquestes dades categòriques en el data set són de tipus string i el model només accepta valors numèrics. Per solucionar això, podem utilitzar diferents tipus de codificació, i per aquest projecte s'ha decidit degut principalment a la senzillesa del seu enteniment i implementació codificar les dades categòriques amb un ***One-hot encoding*** (Gnat, 2021). Aquesta tècnica consisteix en representar amb un vector binari una variable categòrica.

En el cas de la columna *Sex*, com comptem amb valor *female* i *male*, crearem dues columnes més que tindran com a nom *Sex_female* i *Sex_male* respectivament. Llavors, el contingut d'aquesta columna serà el següent: per a cada fila del conjunt, si el client és de sexe masculí, a la columna de sexe femení introduirem 0 i en la del sexe masculí 1. Si fem això per a cada un dels atributs categòrics, ens trobem que el conjunt de dades ha augmentat la seva mida de forma considerable. S'ha de tenir en compte, però, que la columna original de cada un dels atributs categòrics és eliminada del conjunt.

	Age	Credit amount	Duration	Sex_female	Sex_male	Job_0	Job_1	Job_2	Job_3	Housing_free	...	Checking account_unknown	Purpose_business	Purpose_car
0	0.857143	0.050567	0.029412	0	1	0	0	1	0	0	...	0	0	0
1	0.053571	0.313690	0.647059	1	0	0	0	1	0	0	...	0	0	0
2	0.535714	0.101574	0.117647	0	1	0	1	0	0	0	...	1	0	0
3	0.464286	0.419941	0.558824	0	1	0	0	1	0	1	...	0	0	0
4	0.607143	0.254209	0.294118	0	1	0	0	1	0	1	...	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.214286	0.081765	0.117647	1	0	0	1	0	0	0	...	1	0	0
996	0.375000	0.198470	0.382353	0	1	0	0	0	1	0	...	0	0	1
997	0.339286	0.030483	0.117647	0	1	0	0	1	0	0	...	1	0	0
998	0.071429	0.087763	0.602941	0	1	0	0	1	0	1	...	0	0	0
999	0.142857	0.238032	0.602941	0	1	0	0	1	0	0	...	0	0	1

Figura 23: Data set resultant de realitzar una codificació one-hot. Font: elaboració pròpia

En el següent pas es tracta la columna *Risk*, la qual és el nostre target o variable a la que volem donar una predicció. En aquest cas específic, la predicció únicament por obtenir un valor binari: *Good* en cas que el risc sigui acceptable, o bé *bad* en cas que no sigui acceptable. Aquesta columna, com hem explicat anteriorment, no és acceptada per els models ja que no compte amb un valor numèric, i al tractar-se de la variable target, no podem realitzar un one-hot enconding.

Així què, per a poder ser acceptat, simplement es canviarà el valor de *Good* per un 1 i el valor *Bad* per a un 0.

Per finalitzar amb el tractament, per a poder entrenar al model que utilitzarem més endavant, s'ha de crear un subconjunt d'entrenament i un subconjunt de test. El primer subconjunt s'usarà per a entrenar al model a donar una predicció i el subconjunt de test ens permetrà realitzar un test de l'entrenament del model.

Per fer això, mitjançant la funció ***traint_test_split*** crearem un subconjunt d'entrenament i un altre subconjunt de test. Per a aquest cas, realitzarem la distribució de la següent forma: un 75% de les dades seran usades per a l'entrenament, mentre que el restant 25% serà utilitzat per test.

Gràcies a aquests valors podrem entrenar al model i testejar-lo, podent calcular a més un valor de precisió, tant per les dades d'entrenament com per a les dades de test.

4.5. Preparació de l'entorn de desenvolupament i procés d'instal·lació

Donat que Anaconda Navigator permet crear entorns per evitar incompatibilitats entre llibreries, s'ha creat un entorn específic per a utilitzar AIX360. L'entorn es troba disponible al repositori del projecte (March, 2024). A la hora de preparar l'entorn, primer de tot, s'haurà de tenir en compte quin és el mètode que s'utilitzarà ja que, per començar, no tots els mètodes són compatibles amb totes les versions de Python. A la documentació del repositori, es troba la següent taula:

Installation keyword	Explainer(s)	OS	Python version
cofrnet	cofrnet	macOS, Ubuntu, Windows	3.10
contrastive	cem, cem_maf	macOS, Ubuntu, Windows	3.6
dipvae	dipvae	macOS, Ubuntu, Windows	3.10
gce	gce	macOS, Ubuntu, Windows	3.10
imd	imd	macOS, Ubuntu	3.10
lime	lime	macOS, Ubuntu, Windows	3.10
matching	matching	macOS, Ubuntu, Windows	3.10
nncontrastive	nncontrastive	macOS, Ubuntu, Windows	3.10
profwt	profwt	macOS, Ubuntu, Windows	3.6
protodash	protodash	macOS, Ubuntu, Windows	3.10
rbm	brcg, glrm	macOS, Ubuntu, Windows	3.10
rule_induction	ripper	macOS, Ubuntu, Windows	3.10
shap	shap	macOS, Ubuntu, Windows	3.6
ted	ted	macOS, Ubuntu, Windows	3.10
tsice	tsice	macOS, Ubuntu, Windows	3.10
tslime	tslime	macOS, Ubuntu, Windows	3.10
tssaliency	tssaliency	macOS, Ubuntu, Windows	3.10

Figura 24: Taula de compatibilitats dels explicadors amb la versió de Python. Font: (IBM, 2023)

La columna *Explainer* indica el mètode explicatiu i la columna *Python version* indica quina versió de Python es compatible amb cada mètode explicatiu.

Un cop identificat el mètode que volem utilitzar, s'engegarà Anaconda i en el menú situat al lateral esquerra es seleccionarà la opció *environaments*. Un cop polsat, ens apareixerà un llistat amb els entorns virtuals que hi hagi creats a l'ordinador.

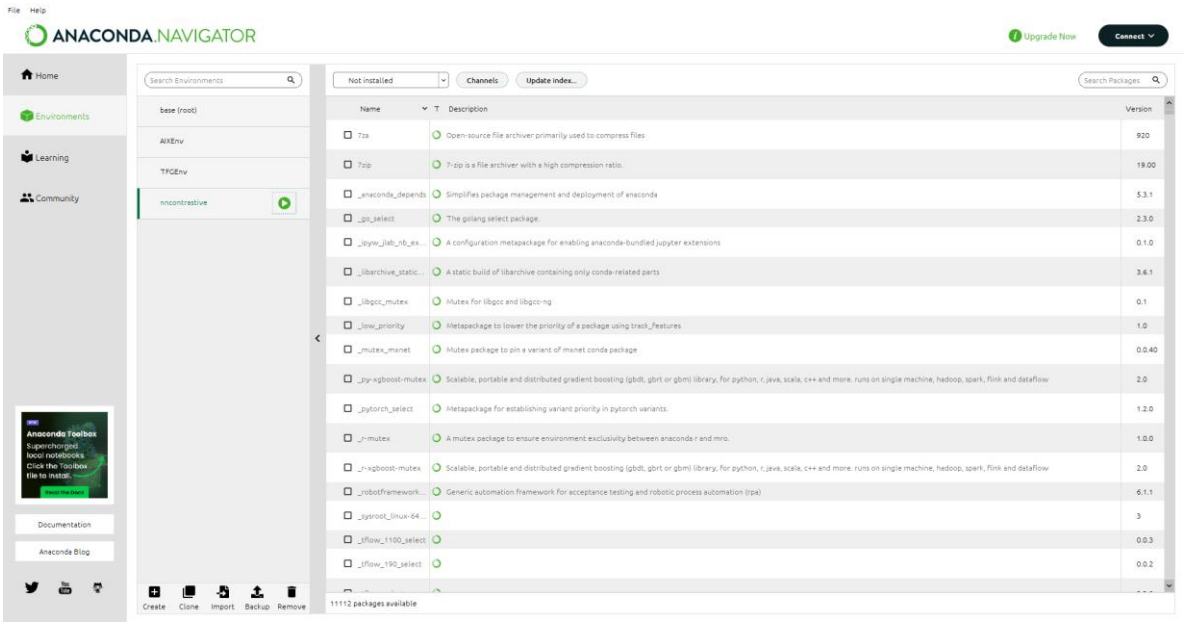

Figura 25: Llistat de entorns virtuals creats. Font: elaboració pròpia

A la part inferior esquerre hi ha una barra d'eines amb 5 botons. Si es polsa el botó *Create*, apareixerà una finestra emergent on s'introduirà primer de tot quin nom rebrà l'entorn en qüestió i posteriorment, un parell de Checkboxs que permetran escollir si volem un entorn de R o Python. Es polsarà a la checkbox de Python i mitjançant el desplegable del costat, s'escollirà la versió especifica de Python d'aquest entorn, el qual haurà de variar segons el mètode escollit.

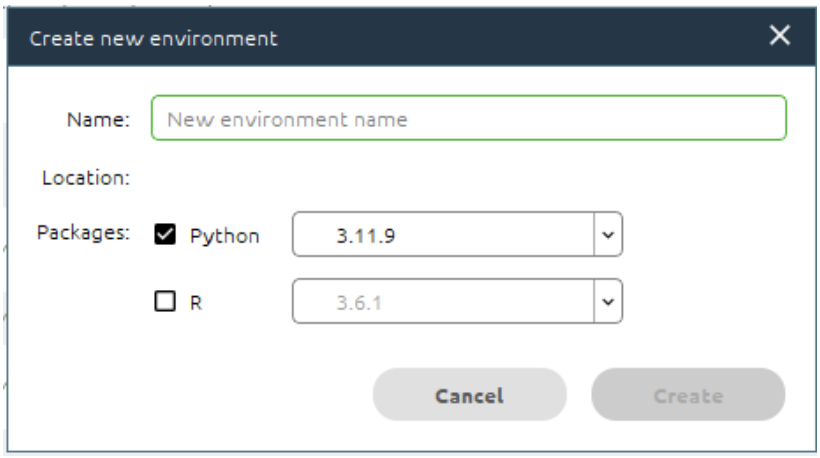

Figura 26: Formulari de creació d'entorn virtual. Font: elaboració pròpia

Un cop emplenats els camps anteriors, es polsarà el botó *Create*. En el moment que l'entorn estigui preparat, ja es podrà començar a afegir les llibreries necessàries. En aquest apartat

explicarem realitzar la instal·lació de les llibreries mitjançant la consola, tal i com s'ha realitzat durant la fase de desenvolupament. Per accedir a una terminal amb un entorn virtual actiu, es retrocedirà al llistat d'entorns virtuals mostrat a la Figura 25 i es polsarà al botó *Play* que hi ha al costat. A la petita finestra emergent que apareixerà es seleccionarà l'opció *Open terminal*. Un cop fet, apareixerà una terminal amb l'entorn virtual actiu i on es podrà començar a instal·lar totes les llibreries, ja sigui amb pip, conda o qualsevol altre instal·lador de paquets.

La primera instal·lació que es realitzarà en el nostre entorn serà, en cas de voler treballar amb Jupyter Notebook, l'entorn complert de Jupyter. Per fer això, s'introduirà a la consola la següent comanda: *pip install jupyter*. Al executar la comanda, s'instal·larà totes les llibreries de Jupyter al nostre entorn.

Un cop s'hagi acabat l'operació, es procedirà a la instal·lació de la llibreria AIX360, que requerirà de més operacions. Primer, es crearà a un directori de treball i haurem de realitzar la clonació del repositori de GitHub. Per fer això, s'introduirà a la mateixa terminal la següent comanda: *git clone https://github.com/Trusted-AI/AIX360*. Al executar la comanda, es descarregarà tot el contingut del repositori en una carpeta anomenada *AIX360, dins d*el directori actual. S'accedirà a la carpeta *AIX360* i posteriorment, s'haurà de buscar a la Figura 24 **¡Error! No se encuentra el origen de la referencia.**quina paraula clau es requereix per a poder instal·lar l'explicador. Per exemple, si es volgués utilitzar l'explicador cofrnet, s'haurà d'utilitzar la paraula clau *cofrnet*. Un cop seleccionada la paraula clau, s'escriurà la següent comanda: *pip install -e .[<paraula_clau>, ...]*. Dins de la llista introduirà la clau de l'explicador que s'utilitzarà. Un cop introduïda la comanda, es realitzarà la instal·lació de totes les llibreries necessàries per als explicadors en el entorn preparat.

Finalment, únicament s'haurà d'instal·lar una llibreria més, **joblib**, una llibreria que ens permet emmagatzemar un model ja entrenat en un fitxer. Per instal·lar-ho, s'haurà d'introduir la següent comanda: *pip install joblib*. Un cop finalitzat aquest últim procés ja es podrà començar a utilitzar la llibreria AIX360.

4.6. Preparació del projecte

Per a preparar el classificador binari en qüestió, es farà ús de un mètode post-hoc, o sigui, el model primer de tot generarà una predicció amb les dades d'entrada, per posteriorment un algorisme a part del model generarà les explicacions contrafactuals.

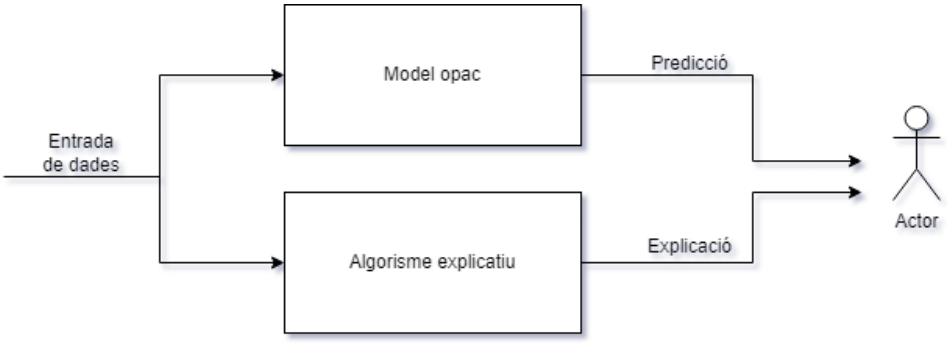

Figura 27: Diagrama de la estructura del classificador binari. Font: elaboració pròpia

S'ha decidit preparar un sistema seguint aquesta tècnica, primer de tot, perquè aquest projecte es centra en un tipus molt concret d'explicació: els contrafactuals. Tenint en compte això, l'objectiu no és implementar un model lo suficientment interpretable, sinó que el classificador sigui capaç de retornar explicacions contrafactuals per a aquelles sol·licituds de crèdits classificades com a denegades. A part, com s'ha exposat a l'apartat 3.1, aquest tipus de tècnica permet utilitzar models amb molt més rendiment, que permet implementar millores del model en treballs futurs.

En el següent apartat, primer de tot, es defineix tot el procés que hi ha hagut darrere de l'elecció del model que genera les prediccions en el nostre classificador binari. Posteriorment, s'exposarà quin ha estat el mètode per generar explicacions escollit i com s'utilitzarà.

4.6.1. Model darrere el sistema

Aquesta preparació, malgrat el projecte no es centra específicament en el Machine Learning, compta amb un model que serà entrenat amb les dades exposades a l'apartat 4.1. Durant el desenvolupament s'han provat diferents opcions de cara a poder desenvolupar un model d'intel·ligent que s'exposen a continuació.

La primera opció que es va rumiar va ser la possibilitat de desenvolupar una xarxa neuronal, definint totes les seves capes. Per fer-ho, es va fer ús de la llibreria *Keras*, una llibreria de Python que simplifica la implementació de models. El plantejament era crear un model usant la classe *Sequential*, que consisteix en una pila lineal de capes en un model. Un cop creat l'objecte, el següent pas és determinar quines capes formaran part del model.

El criteri que es va utilitzar per determinar aquest model va ser el següent: donat que el problema que volem resoldre no és considera un problema de gran complexitat, ni tampoc haurem de manejar un gran volum de dades, el model en qüestió no requeria d'una gran quantitat de capes, ni cada capa podia contenir una gran quantitat de neurones. Això es deu principalment a que l'ús de més capes i neurones de les necessàries pot generar problemes de overfitting.

Per això, es va plantejar un model de xarxa neuronal molt simple, el qual es va implementar amb la llibreria *Keras*, amb dues capes de neurones: una primera capa d'entrada de les dades amb funció d'activació *ReLU* i una segona capa de sortida, amb una sola neurona per a retornar únicament un valor i amb funció d'activació *Sigmoid*, la qual té com a característica principal que retorna valors compresos entre 0 i 1, el qual s'interpreta com a un valor de percentatge de probabilitat.

Malgrat plantejar crear una xarxa neuronal simple, els resultats no eren els esperats, ja que obteníem uns valors de precisió molt baixos. Degut a això, i tenint en compte la gran complexitat al darrere de desenvolupar les capes del model, a més que el projecte es centra en la generació de les explicacions i no en el model de Machine Learning, es va descartar l'opció de desenvolupar un model propi.

A partir de prendre la decisió anterior, es va decidir fer ús de un model ja existent. Aquest model ha estat un *DecisionTreeClassifier* de la llibreria Scikit Learn el qual consisteix en un tipus de model no paramètric molt usa per classificacions i regressions (Sklearn.Tree.DecisionTreeClassifier, 2024). S'ha escollit aquest model ja que el seu funcionament és simple d'entendre, a més que el model en qüestió es troba implementat i es pot instanciar únicament important la llibreria. A més, el propi model permet retocar els paràmetres (com la profunditat màxima de l'arbre, el número màxim de nodes fulla, ...).

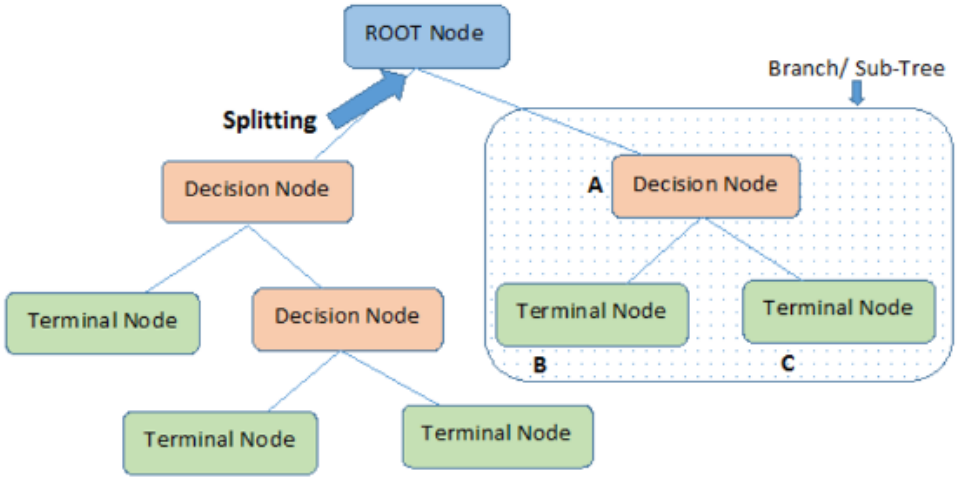

Figura 28: Representació estructural de un arbre de decisions. Font: (Saini, 2024)

Com veiem a la il·lustració anterior, veiem que l'arbre comença amb un node origen el qual va generant un ordre jeràrquic de arbres inferiors. Com més profund sigui l'arbre, més complexa serà la decisió i millor s'ajustarà a les dades.

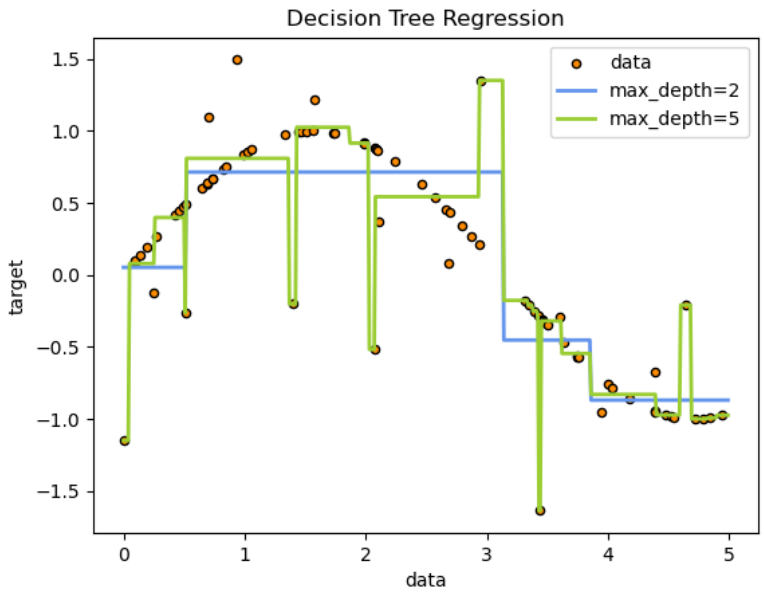

Figura 29: ajust del model a les dades segons la profunditat de l'arbre. Font: (Decision Tree, 2024)

A la gràfica anterior, exposada a la web de la llibreria de Scikit Learn (*Decision Tree*, 2024), s'observa un pla amb les dades i, mostrat en color bau, l'ajust d'un arbre de decisions amb una profunditat màxima de dos i mostrat en color ver l'ajust d'un altre arbre de decisions amb una profunditat màxima de 5. L'arbre binari representat en color blau es troba relativament ajustat a les dades, però amb un ample rang de millora, mentre que l'arbre representat en color verd, al tenir una profunditat màxima més alta, s'ajusta de forma més pròxima a les dades, generant prediccions més ajustades.

El problema recau en el moment que es pot generar *overfitting,* o sigui, que el model s'ajusti molt bé a les dades d'entrament i en el moment que comença a generar prediccions amb dades que no estan presents a les dades d'entrenament no obté bons resultats. Aquest fenomen va succeir durant la preparació del classificador, ja que en una primera instància, el model retornava una precisió de 1'0 per les dades d'entrenament i un precisió de 0'56 per a les dades de test. Com s'observa amb aquests resultats, el model s'ha ajustat massa bé a les dades d'entrenament i per això retorna un precisió tan alta, però un cop posem a prova el model amb les dades de test, aquest no és capaç de generar prediccions encertades, fent que precisió del model baixi dràsticament.

Donat que disposem de un model d'arbre de decisions, les tècniques aplicades per reduir el overfitting han estat les següents:

- **Aplicar una profunditat màxima:** com s'ha exposat anteriorment, aquest tipus de model genera diferents subnivells formant un arbre. Com més profunditat té aquest arbre, més complexes i ajustades són les prediccions. En aquest cas, tenint en compte que tampoc es compta amb una gran quantitat de dades, la profunditat de l'arbre no podrà ser massa gran, per la qual cosa es limitarà la profunditat màxima de l'arbre a 10.

- **Aplicar un número màxim de nodes fulla:** per entendre bé aquest punt, primer es definirà un node fulla. Aquest tipus de nodes són aquells nodes que no tenen cap mena de descendència, o sigui, d'ells no apareix un altre arbre. Si s'ajusta aquest valor màxim, s'aconsegueix, de la mateixa forma que la profunditat màxima, simplificar l'arbre per evitar que aquest creixi massa i produeixi una sobre ajust del model.

Instància del model	Accuracy d'entrenament	Accuracy de test
DecissionTreeClassifier()	1.0	0.56
DecissionTreeClassifier(max_depth=10 max_leaf_nodes=10)	0.74	0.7

Taula 4: taula resum de la diferència de rendiment entre les diferents implementacions del model. Font: elaboració pròpia

A partir d'aplicar aquestes dues tècniques, s'ha aconseguit reduir el overfitting: els valors que s'obtenen ara al calcular la precisió amb les dades d'entrenament és d'aproximadament el 0.74, mentre que la precisió amb les dades de test es situa aproximadament al 0.70. S'ha reduït la precisió per l'entrenament, però s'ha aconseguit a la vegada un increment de la precisió de test.

De totes formes, malgrat amb aquesta implementació de l'arbre de decisions s'ha millorat el rendiment respecte l'arbre de decisions inicial, segueixen sent un valors bastant baixos per considerar-se acceptables. Un rendiment òptim per a un classificador binari ha de situar-se al voltant del 0.8 i 0.85, tan amb les dades d'entrenament com a les dades de test. Aquest problema de rendiment baix pot ser causat per certs motius.

El primer d'aquests motius pot ser **el tipus de model utilitzat**. Com hem mencionat anteriorment a l'apartat de teoria, un model d'arbre de decisions es considera un model interpretable degut principalment a l'estructura simple del model. Degut a aquesta senzillesa, és molt probable que el model no rendeixi del tot bé.

Un altre motiu per el qual és possible que el model retorni uns valors de precisió molt baixos pot ser degut a **la baixa densitat de dades** utilitzades per a l'entrenament. Com hem explicat en el apartat 4.1, el nostre data set únicament inclou 1000 instàncies de clients, un volum que es podria considerar molt baix.

4.6.2. Mètode per generar explicacions

Un cop escollit el model, el següent pas serà escollir quin serà el mètode que usarem per a generar les explicacions.

En un inici, el mètode que es va escollir va ser l'explicador *Contrastive Explainer* en mode "Pertinent Negatives" (Dhurandhar et al., 2018), però durant el desenvolupament van sorgir problemes amb aquest mètode ja que el contrafactual que generava sempre retornava tots els atributs amb valor 0. Es va investigar quins podrien ser els motius darrere aquests resultats, però malgrat llegir la documentació de la llibreria i cercar informació en altres fonts, no es va aconseguir que els contrafactuals generats fossin acceptables, motiu que va portar a cercar un altre explicador capaç de generar contrafactuals.

Degut a la falta de resultats i a l'alta complexitat de la solució, es va cercar un altre explicador que també pot generar contrafactuals inclòs a la llibreria AIX360. Aquest explicador és el *Nearest Neighbor Contrastive Explainer*. Aquest algorisme genera una representació de baixa dimensionalitat per a cada punt de dades, amb l'objectiu de cercar veïns a aquests punts. Llavors, donada la entrada de dades a aquest mètode, l'algorisme cercarà aquells veïns més pròxims a les dades introduïdes que tinguin una etiqueta de classe diferent. En el context d'aquest projecte, en el cas de introduir en el mètode una instància de un client al que se li ha denegat el crèdit, llavors l'explicador retornarà la sol·licitud de crèdit classificada com a concedida més propera a aquesta instància.

A més, aquest mètode té l'avantatge que permet definir a la crida del mètode el número de veïns a generar. Llavors, configurant aquesta variable, es generen tres possibles canvis per a un client que se li ha denegat un crèdit.

Tenint en compte que ara es disposa de una instància de un client a la que se li ha denegat la sol·licitud de crèdit i un o varis veïns els quals es classifiquen amb un risc positiu, però amb la diferència mínima, ara cal mostrar les dades de forma clara. Per fer això, s'ha creat un codi que permet crear un dataframe amb tres columnes: la primera columna indica la instància original, o sigui, les dades que s'han introduït per generar l'explicació, la segona columna que indica el veí més proper generat per l'explicador i la tercera columna que ens indica la diferència entre la primera i la segona columna. A més, totes aquelles columnes que pateixin alguna variació es mostraran de color groc.

	Original value	Predicted value	Difference between instances
Age	25.000000	25.000000	0.000000
Credit amount	1295.000000	1600.000000	305.000000
Duration	12.000000	10.000000	-2.000000
Sex_female	1.000000	1.000000	0.000000
Sex_male	0.000000	0.000000	0.000000
Job_0	0.000000	0.000000	0.000000
Job_1	0.000000	0.000000	0.000000
Job_2	1.000000	1.000000	0.000000
Job_3	0.000000	0.000000	0.000000
Housing_free	0.000000	0.000000	0.000000
Housing_own	0.000000	0.000000	0.000000
Housing_rent	1.000000	1.000000	0.000000
Saving accounts_little	1.000000	1.000000	0.000000
Saving accounts_moderate	0.000000	0.000000	0.000000
Saving accounts_quite rich	0.000000	0.000000	0.000000
Saving accounts_rich	0.000000	0.000000	0.000000
Saving accounts_unknown	0.000000	0.000000	0.000000
Checking account_little	0.000000	0.000000	0.000000
Checking account_moderate	1.000000	1.000000	0.000000
Checking account_rich	0.000000	0.000000	0.000000
Checking account_unknown	0.000000	0.000000	0.000000
Purpose_business	0.000000	0.000000	0.000000
Purpose_car	1.000000	1.000000	0.000000
Purpose_domestic appliances	0.000000	0.000000	0.000000
Purpose_education	0.000000	0.000000	0.000000
Purpose_furniture/equipment	0.000000	0.000000	0.000000
Purpose_radio/TV	0.000000	0.000000	0.000000
Purpose_repairs	0.000000	0.000000	0.000000
Purpose_vacation/others	0.000000	0.000000	0.000000
Risk	Bad	Good	1.000000

Il·lustració 1: dataframe amb el valor original, el veí creat i la seva diferència. Font: elaboració pròpia

Per posar un exemple del resultat que s'obté al final de l'execució, la columna *Original value* conté la instància original del client, la columna *Predicted value* conté el contrafactual (en aquest

cas únicament s'ha generat 1 contrafactual) i la ultima columna *Difference between instances* indica la diferència entre les dues primeres columnes. En aquest cas, el plantejament que ha realitzat l'explicador és el següent: per a que la sol·licitud del crèdit passi de ser negatiu a positiu el client hauria d'augmentar de forma molt lleugera el crèdit sol·licitat a la vegada que redueix en dos mesos el temps de retorn del crèdit.

4.7. Definició del disseny experimental

Com s'ha mencionat anteriorment, el principal objectiu del projecte és realitzar una prova d'avaluació amb usuaris on volem respondre la següent pregunta: quin format de representació de les explicacions contrafactuals millora la **comprensió** i **satisfacció** dels usuaris? i que es concreta encara més en la comparació de dos formats, en format tabular o bé en text?

Per a la pregunta proposada es planteja la següent hipòtesi:

- El format textual millora la comprensió i la satisfacció dels usuaris envers del format tabular.

Es planteja aquesta hipòtesi ja que altres recerques han trobat que les explicacions textuals milloren la comprensió dels usuaris (Wang & Yin, 2021).

El primer pas per poder realitzar aquesta prova serà definir el disseny experimental. Violet Turri (Violet Turri, 2022) argumenta que una de les limitacions de la intel·ligència artificial explicable és la manca de consens a l'hora de testejar les explicacions. Per això, en aquest projecte es planteja un disseny experimental basat en el disseny exposat per Clara Bove en el seu article *Investing the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study* (Bove et al., 2023).

El disseny experimental cerca avaluar quin format de contrafactuals millora la comprensió i satisfacció dels usuaris. Per això, es proposa un disseny experimental *between subjects*, ja que cada participant de l'estudi se li mostrarà únicament un format. S'ha escollit aquest tipus de disseny ja que, com indica Julia Simkus (Simkus, 2023), malgrat aquest tipus de disseny requereix d'un ventall més gran de usuaris, també permet reduir l'efecte *carryover,* que consisteix en que les proves que s'han realitzat anteriorment puguin afectar als resultats de les proves posteriors. En el context d'aquest projecte, aquest efecte pot succeir ja que, en cas d'avaluar amb el mateix participant els dos formats, si en un inici es mostra un format i el participant entén a la perfecció l'explicació, la prova del següent format es pot veure alterada degut a la comprensió del primer format.

4.7.1. Criteris d'inclusió i exclusió

En aquest apartat s'exposa quins seran els criteris d'inclusió i exclusió dels participants de la prova d'avaluació, o sigui, les característiques que ha de tenir un participant per a no descartar-se la seva participació.

Els criteris d'inclusió definits per a aquesta prova d'avaluació són els següents:

- El participant ha de tenir una edat mínima de 18 anys.
- El participant ha d'haver sol·licitat un crèdit en alguna ocasió.

Els criteris d'exclusió definits per a aquesta prova d'avaluació són els següents:

- El participant no pot tenir coneixements avançats respecte al procés per a determinar si un crèdit és denegat o no.
- El participant no pot tenir tampoc coneixements avançats d'intel·ligència artificial ni de mètodes per generar explicacions.
- El participant no pot tenir cap coneixement respecte a què és una explicació contrafactual.

El total de participants de l'estudi és de 30 usuaris.

4.7.2. Recursos utilitzats

Per a poder realitzar la prova d'avaluació, s'han creat un seguit de recursos que seran necessaris per realitzar l'avaluació.

El primer recurs i el més important serà els **contrafactuals** que s'exposarà als participants. Per generar-los, primer s'ha generat un total de 6 situacions en les que es presenta una sol·licitud i aquesta es classifica com a denegada. Aquestes situacions consisteixen en una persona de sexe femení i un altre de sexe masculí per a cada una de les franges d'edat generades. Aquestes franges d'edat s'han distribuït en:

- Usuaris dels 19 als 28 anys
- Usuaris dels 29 als 38 anys
- Usuaris dels 39 anys fins als 75 anys

S'han definit aquestes franges d'edat per aconseguir una distribució de les sol·licituds de crèdit exposades a la Figura 9 equivalent, o sigui, que per a cada franja d'edat hi hagi un número similar de persones.

Per a cada situació s'ha generat un total de 3 contrafactuals diferents, únicament amb canvis en la duració del crèdit i en la quantitat de crèdit sol·licitat, amb la finalitat d'augmentar la potència de la prova d'avaluació. De cada un dels contrafactuals, a més, es genera una versió tabular i una versió textual.

L'ús que se li donarà a aquestes situacions i contrafactuals serà el següent: es seleccionarà per a cada participant una de les 6 situacions exposades anteriorment. En el inici de la prova, es realitzarà una introducció en la que s'exposarà al participant la situació escollida, simulant que és el propi participant qui demana el crèdit. Un cop feta aquesta introducció i després de resoldre

possibles dubtes del participant, se li mostrarà els contrafactuals en un dels dos formats que cerquem avaluar.

El següent recurs que s'utilitzarà serà un **formulari de comprensió**, en el que es cercarà avaluar què ha entès l'usuari amb els contrafactuals, sigui en el format que sigui. Primer de tot, l'usuari haurà de marcar amb una creu quin format de contrafactual se li ha exposat: tabular o textual. Posteriorment, el formulari exposa un seguit d'afirmacions i per a cada una, el participant haurà de marcar si està d'acord, si no ho està o bé no ho té clar.

Les afirmacions que es proposen en aquest formulari són les següents:

1. La informació que es mostra indica les variacions que es poden fer per a que es concedeixi el crèdit.
2. El número màxim de característiques de la sol·licitud que s'han de canviar no és mai superior a dos.
3. El motiu de la sol·licitud del crèdit pot canviar l'estat de la sol·licitud.
4. Tots els contrafactuals indiquen que la quantitat de crèdit ha d'augmentar per a que l'estat de la sol·licitud canviï.
5. Com a mínim es planteja un possible canvi en la duració del crèdit.
6. El lloc del treball pot causar que l'estat del crèdit canviï.

Finalment, el formulari compta amb una pregunta oberta, la qual el participant haurà de respondre de forma breu i concisa. La pregunta a respondre és la següent:

- Quins possibles canvis es proposen per a que el crèdit canviï el seu estat de denegat a concedit?

Aquest formulari tindrà el següent format de puntuació: per a cada afirmació hi ha una resposta correcta predeterminada. Aquestes respostes predeterminades són les següents:

1. Estic d'acord
2. Estic d'acord
3. No estic d'acord
4. No estic d'acord
5. Estic d'acord
6. No estic d'acord

Per a cada resposta del participant que coincideixi amb aquesta resposta predeterminada es sumarà un punt al total de puntuació. A més, si el participant a la resposta de la pregunta oberta indica tots els canvis que es proposen en les explicacions contrafactuals es sumarà un últim punt, podent obtenir finalment un resultat situat entre 0 i 7. Es planteja aquest format de puntuació per a determinar si el participant ha entès les explicacions que està rebent.

L'últim recurs que utilitzarem serà un **qüestionari de satisfacció**, adaptat de Robert R. Hoffman (Hoffman et al., 2018). L'adaptació que s'ha realitzat ha estat, primer de tot, adaptar les afirmacions al context d'aquesta prova d'avaluació, encara que les afirmacions segueixen

mantenint la mateixa base. Després, s'ha adaptat el format de puntuació. Hoffman (Hoffman et al., 2018) per a cada afirmació únicament permet tres possibles puntuacions: *Essential*, *Useful but not essential* i *Not necessary*. En aquesta adaptació, s'ha canviat aquest format per una escala Likert amb valors de 1 a 6, ja que Leonard J.Simms (Simms et al., 2019) demostra que les preguntes amb respostes de 6 punts és un format raonable per a estudis psicològics.

En aquest formulari en qüestió, primer de tot el participant haurà d'indicar amb una creu quin tipus de format se li ha mostrat. Posteriorment s'exposen les afirmacions següents:

- Les explicacions rebudes m'han resultat fàcils de comprendre.
- Les explicacions rebudes m'han resultat satisfactòries.
- Les explicacions rebudes són suficientment detallades.
- Les explicacions rebudes no compten amb informació no rellevant.
- Les explicacions rebudes són molt complertes.
- Les explicacions rebudes són útils per a entendre el motiu de la denegació del crèdit.
- Les explicacions rebudes per obtenir el crèdit són precises.
- Les explicacions rebudes per obtenir el crèdit són fiables.

El format de puntuació d'aquest formulari serà el següent: com s'ha mencionat anteriorment, una de les adaptacions que s'ha realitzat és, per a cada afirmació, aplicar un valor seguint una escala Likert de 1 a 6. Llavors, per extreure una puntuació general de la satisfacció del formulari, es calcularà una mitjana dels valors obtinguts en cada afirmació, obtenint una puntuació final entre 1 i 6. Gràcies a aquesta puntuació, podrem generar un valor de satisfacció general, que ens permetrà més endavant realitzar una comparació entre la satisfacció dels dos formats.

4.7.3. Procediment de la prova d'avaluació

La prova d'avaluació proposada té el següent procediment: l'estudi es realitzarà en un aula amb un moderador present, que es trobarà present en tot moment de la prova per resoldre els dubtes en cas que el participant en tingui. En el moment que el participant indiqui que està preparat per començar la prova, el moderador realitzarà una explicació detallada de quins passos es seguiran durant la prova. Posteriorment, es resoldran els dubtes que puguin sorgir al participant i un cop resolts, començarà la prova. A més, també començarà a gravar-se la prova.

El participant se li exposarà una de les sis situacions en les que es presenta una sol·licitud de crèdit i aquesta es classifica com denegada. Posteriorment, s'exposarà al participant, en format textual o format tabular, les explicacions contrafactuals de la situació exposada. En el mateix moment que el participant rep les explicacions contrafactuals, rebrà a més el formulari de comprensió i el formulari de satisfacció. Primer omplirà el formulari de comprensió, indicant quin format se li ha mostrat durant la prova d'avaluació i responent a les afirmacions i la pregunta oberta proposades.

Un cop hagi finalitzat el formulari de comprensió, aquest formulari serà recollit per el moderador i les explicacions contrafactuals seran retirades de la vista del participant. En aquest moment, el

participant omplirà el formulari de satisfacció, primer indicant quin format se li ha mostrat durant la prova d'avaluació i donant un valor d'acord a les afirmacions proposades en el formulari.

Un cop el formulari de satisfacció hagi estat emplenat, la prova d'avaluació haurà finalitzat.

4.7.4. Càlcul i anàlisi dels resultats obtinguts

Un cop finalitzada la prova d'avaluació amb els usuaris, començarà l'etapa de càlcul i anàlisi dels resultats obtinguts dels formularis. En aquesta etapa es disposarà de la següent informació:

a. 15 formularis de comprensió per al format textual
b. 15 formularis de satisfacció per al format textual
c. 15 formularis de comprensió per al format tabular
d. 15 formularis de satisfacció per al format tabular

Per respondre a la pregunta proposada per aquest estudi, primer extraurem una avaluació de la comprensió de cada format. Per fer això, es calcularà una puntuació mitjana dels formularis de comprensió, tan per el format tabular com per el format textual. Un cop extretes aquestes dues mitjanes, es realitzarà una comparació dels resultats obtinguts per comprovar quin dels dos formats ha millorat més la comprensió dels participants.

Un cop haguem extret la puntuació mitjana de comprensió per el format tabular i textual, es realitzarà un procediment similar per a extreure una avaluació de la satisfacció generada per el format tabular i el format textual. Es calcularà una puntuació mitjana dels formularis de satisfacció del format tabular i una puntuació mitjana del formularis de satisfacció del format textual. Un cop extretes aquestes dues mitjanes, es realitzarà una comparació dels resultats obtinguts per comprovar quin dels dos formats ha generat més satisfacció entre els participants.

Com s'ha exposat a l'apartat anterior, la prova d'avaluació es gravarà. Aquesta gravació es realitza amb la finalitat d'enregistrar quines reaccions tenen els participants a la hora de rebre les explicacions, alhora que s'enregistraran les frases i preguntes que realitzarà el participant durant la prova. D'aquesta forma, podem establir quines solen ser les reaccions inicials al rebre les explicacions en cada format, a part de constatar quins són els dubtes més preguntats per a cada format. Aquests resultats es corroboraran amb els resultats obtinguts dels formularis, comprovant si les reaccions que puguin fer els participants i les preguntes realitzades coincideixen amb la comprensió i satisfacció obtinguda dels formularis.

5. Conclusions

Com a conclusió del projecte, es pot dir que aquests models explicables prometen millorar molt la relació entre les persones i la intel·ligència artificial. Cada dia els model es troben més presents en el nostre entorn, i justament per això s'han d'implementar noves tècniques i algorismes que permetin, per una banda, als usuaris entendre el resultat que ha rebut per a millorar la confiança en els sistemes, i per un altre banda als desenvolupadors i experts a entendre que està fent el model per a poder millorar-lo progressivament.

Durant el desenvolupament del projecte, primer de tot, s'ha realitzat una pinzellada de teoria d'aquests models explicables. En ell s'ha analitzat primer de tot què és un model d'intel·ligència artificial explicable, mencionant els interessos que hi ha al darrere de l'estudi d'aquests algorismes. També s'ha exposat què és una explicació i s'ha centrat el projecte en un tipus molt concret d'explicacions: les explicacions contrafactuals. Aquest punt ha resultat molt útil per a desenvolupar l'apartat pràctic del projecte, però també ha ajudat a augmentar els coneixements en intel·ligència artificial.

Per al desenvolupament pràctic, s'ha treballat amb el llenguatge Python i amb la llibreria AIX360 per preparar un classificador binari per classificar sol·licituds de crèdit com a denegades o concedides. A més, també s'ha preparat un mètode explicatiu que genera explicacions contrafactuals per a aquelles sol·licituds classificades com a denegades. Aquest apartat ha resultat molt útil per entendre l'explicabilitat post hoc. S'ha fet un disseny d'avaluació amb usuaris en el que es cerca avaluar quin format millora la comprensió i satisfacció dels usuaris.

Els objectius plantejats per a l'apartat teòric han estat assolits, però s'han d'exposar els problemes que han sorgit durant el desenvolupament de l'apartat pràctic. La preparació del classificador es va allargar més temps del desitjat, factor que ha portat a que la prova d'avaluació no es pogués realitzar abans de lliurament d'aquesta memòria. Per compensar aquest retard amb els resultats de la prova d'avaluació, es planteja realitzar la prova d'avaluació i mostrar els resultats a la presentació final.

5.1. Treball futur

Durant el desenvolupament del projecte, especialment l'apartat pràctic, a mesura que avançava el projecte, han aparegut possibles millores que, degut principalment a falta de temps no s'han dut a terme. Aquestes millores s'exposen a continuació.

1. Canvi de conjunt de dades: en aquest projecte s'ha utilitzat com a dades d'entrenament un conjunt de dades amb una densitat molt petita (1000 instàncies). Com a possible millora es podria cercar un nou data set que permetés millorar la fase d'entrenament del model.

2. Millora de rendiment: malgrat que aquest projecte no es centrava en el rendiment del model, una proposta de millora seria millorar el rendiment del model per a generar millors prediccions de classificació. Per a realitzar aquesta millora, es plantegen dues opcions:

a. Dissenyar i implementar un model: aquesta proposta proposa plantejar des de 0 un model que s'ajusti al problema que volem resoldre. Un possible opció seria implementar una xarxa neuronal.

b. Perfeccionar la preparació de l'arbre de decisions: l'arbre de decisions compte amb un gran número de paràmetres de configuració que poden millorar la precisió del model. Es pot fer un estudi de tots els paràmetres i com afecta al rendiment d'aquest per realitzar una configuració adequada.

Donat que la prova amb usuaris no s'ha pogut realitzar degut al endarreriment produït durant la preparació del classificador, en aquest apartat no es defineixen millores per al disseny experimental, malgrat que segur apareixen propostes de millora per al disseny experimental un cop que es realitzi la prova d'avaluació.

6. Annexos

En aquest annex s'inclourà primer de tot un glossari amb paraules clau i la seva definició i tots els recursos usats.

6.1. Glossari

- AIX360: llibreria originalment implementada per IBM utilitzada en aquest projecte per a preparar el classificador.
- Anaconda Navigator: entorn de desenvolupament per Python i R.
- Classificador: model que rep una instància de dades i segons les seves característiques la classifica en classes. En cas de tenir únicament dos possibles classes, es tracta de un classificador binari.
- Conjunt d'entrenament: subconjunt de dades del data set original que s'utilitzaran per a entrenar al model.
- Conjunt de test: subconjunt de dades del data set original que s'utilitzarà per a testejar el model.
- Explicabilitat post hoc: conjunt de tècniques i algorismes que permeten generar una explicació de un model posteriorment a que aquest hagi generat una predicció.
- Explicació contrafactual: tipus d'explicació que donat un esdeveniment, indica què hauria de canviar per a que aquest esdeveniment també canviés.
- Intel·ligència artificial explicable: camp de la intel·ligència artificial que fa recerca de tècniques per a que els resultats de un model siguin més comprensibles.
- Model d'arbre de decisions: model molt usat en classificació com regressió. Per a cada atribut genera un subarbre i una subdivisió de les dades, fins que arriba a un node fulla, que resulta en la predicció.
- Model opac: tipus de model d'intel·ligència artificial el qual destaca per el seu rendiment però també per ser complicades d'entendre, degut principalment a la seva estructura complexa.
- Model transparent: tipus de model que destaca per permetre entendre de forma més senzilla les decisions que pren. Tenen una estructura molt més simple i fàcil de seguir de forma lògica.
- Overfitting: fenomen que es pot donar en el moment que un model durant el seu entrenament s'ajusta massa bé a les dades d'entrenament, reduint de forma considerable el seu rendiment amb dades no presents en el conjunt d'entrenament.
- Precisió: valor que permet avaluar la precisió d'encerts que produeix un model a la hora de generar prediccions.
- Prova d'avaluació: prova en la que es cerca donar una resposta a una pregunta o esclarir una hipòtesis. La prova es pot realitzar amb usuaris o bé sense usuaris.

6.2. Contrafactuals utilitzats per a la prova amb usuaris

Per a l'avaluació amb usuaris, s'ha generat 6 situacions en la que un client sol·licita un crèdit i aquest es classifica com a denegat. Per a cada una, es genera un total de 3 contrafactuals.

Primer usuari: [[23, "female", "2", "own", "little", "little", 900, 18, "car"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	23	23	23	23
Credit amount	900	900	750	800
Duration	18	8	6	7
Sex female	1	1	1	1
Sex male	0	0	0	0
Job 0	0	0	0	0
Job 1	0	0	0	0
Job 2	1	1	1	1
Job 3	0	0	0	0
Housing free	0	0	0	0
Housing own	1	1	1	1
Housing rent	0	0	0	0
Saving accounts little	1	1	1	1
Saving accounts moderate	0	0	0	0
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	0	0	0	0
Checking account little	1	1	1	1
Checking account moderate	0	0	0	0
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	0	0	0	0
Purpose car	1	1	1	1
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	0	0	0	0
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation	0	0	0	0
Risk	0	1	1	1

Taula 5: usuari 1 amb tres contrafactuals. Font: elaboració pròpia

Segon usuari: [[26, "male", "0", "free", "moderate", "little", 800, 12, "car"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	26	26	26	26
Credit amount	800	800	650	750
Duration	12	8	5	6
Sex female	0	0	0	0
Sex male	1	1	1	1
Job 0	1	1	1	1
Job 1	0	0	0	0
Job 2	0	0	0	0
Job 3	0	0	0	0
Housing free	1	1	1	1
Housing own	0	0	0	0
Housing rent	0	0	0	0
Saving accounts little	0	0	0	0
Saving accounts moderate	1	1	1	1
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	0	0	0	0
Checking account little	1	1	1	1
Checking account moderate	0	0	0	0
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	0	0	0	0
Purpose car	1	1	1	1
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	0	0	0	0
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation	0	0	0	0
Risk	0	1	1	1

Taula 6: usuari 2 amb tres contrafactuals. Font: elaboració pròpia

Tercer usuari: [[29, "female", "1", "free", "little", "little", 9000, 10, "education"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	35	35	35	35
Credit amount	3000	3000	2600	2850
Duration	27	21	24	26
Sex female	1	1	1	1
Sex male	0	0	0	0
Job 0	0	0	0	0
Job 1	0	0	0	0
Job 2	1	1	1	1
Job 3	0	0	0	0
Housing free	0	0	0	0
Housing own	0	0	0	0
Housing rent	1	1	1	1
Saving accounts little	1	1	1	1
Saving accounts moderate	0	0	0	0
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	0	0	0	0
Checking account little	0	0	0	0
Checking account moderate	1	1	1	1
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	0	0	0	0
Purpose car	0	0	0	0
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	1	1	1	1
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation	0	0	0	0
Risk	0	1	1	1

Taula 7: usuari 3 amb tres contrafactuals. Font: elaboració pròpia

Quart usuari: [[37, "male", "1", "free", "little", "moderate", 7500, 27, "business"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	37	37	37	37
Credit amount	7500	7000	7500	7350
Duration	27	22	25	23
Sex female	0	0	0	0
Sex male	1	1	1	1
Job 0	0	0	0	0
Job 1	1	1	1	1
Job 2	0	0	0	0
Job 3	0	0	0	0
Housing free	1	1	1	1
Housing own	0	0	0	0
Housing rent	0	0	0	0
Saving accounts little	1	1	1	1
Saving accounts moderate	0	0	0	0
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	0	0	0	0
Checking account little	0	0	0	0
Checking account moderate	1	1	1	1
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	1	1	1	1
Purpose car	0	0	0	0
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	0	0	0	0
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation	0	0	0	0
Risk	0	1	1	1

Taula 8: usuari 4 amb tres contrafactuals. Font: elaboració pròpia

Cinquè usuari: [[43, "female", "1", "rent", "little", "little", 3000, 29, "vacation/others"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	43	43	43	43
Credit amount	3000	2800	2675	2740
Duration	29	26	23	25
Sex female	0	0	0	0
Sex male	1	1	1	1
Job 0	0	0	0	0
Job 1	1	1	1	1
Job 2	0	0	0	0
Job 3	0	0	0	0
Housing free	0	0	0	0
Housing own	0	0	0	0
Housing rent	1	1	1	1
Saving accounts little	1	1	1	1
Saving accounts moderate	0	0	0	0
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	0	0	0	0
Checking account little	1	1	1	1
Checking account moderate	0	0	0	0
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	0	0	0	0
Purpose car	0	0	0	0
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	0	0	0	0
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation/others	1	1	1	1
Risk	0	1	1	1

Taula 9: usuari 5 amb tres contrafactuals. Font: elaboració pròpia

Sisè usuari: [[60, "male", "3", "own", "unknown", "moderate", 40000, 25, "car"]]

Features	Original	Contraf. 1	Contraf. 2	Contraf. 3
AGE	60	60	60	60
Credit amount	40000	34000	36500	38000
Duration	25	27	28	30
Sex female	0	0	0	0
Sex male	1	1	1	1
Job 0	0	0	0	0
Job 1	0	0	0	0
Job 2	0	0	0	0
Job 3	1	1	1	1
Housing free	0	0	0	0
Housing own	1	1	1	1
Housing rent	0	0	0	0
Saving accounts little	0	0	0	0
Saving accounts moderate	0	0	0	0
Saving accounts quite rich	0	0	0	0
Saving accounts rich	0	0	0	0
Saving accounts unknown	1	1	1	1
Checking account little	0	0	0	0
Checking account moderate	1	1	1	1
Checking account rich	0	0	0	0
Checking account unknown	0	0	0	0
Purpose business	0	0	0	0
Purpose car	1	1	1	1
Purpose domèstic appliances	0	0	0	0
Purpose education	0	0	0	0
Purpose furniture/equipment	0	0	0	0
Purpose radio/TV	0	0	0	0
Purpose repairs	0	0	0	0
Purpose vacation	0	0	0	0
Risk	0	1	1	1

Taula 10: usuari 6 amb tres contrafactuals. Font: elaboració pròpia

6.3. Guió de presentació a l'usuari

En aquest annex, es mostrarà un exemple del guió amb el que introduiríem a un participant a l'escenari experimental. Primer de tot, es realitzaria una breu presentació on exposem què volem cercar amb l'estudi i que es farà. Aquesta presentació serà la següent:

"Bon dia.

Soc el Gerard March, moderador d'aquesta prova d'avaluació. Abans de començar, però, m'agradaria explicar què volem avaluar amb aquesta prova i posteriorment quin serà el procediment a seguir. Aquesta prova d'avaluació cerca comprovar quin format millora més la comprensió i satisfacció dels usuaris a la hora de rebre una explicació generada per un sistema d'intel·ligència artificial: el format tabular o format textual. Per fer-ho, es mostrarà una situació de una persona que demana un crèdit a un banc, però que degut a les característiques d'aquesta sol·licitud, el banc en qüestió la denega. Posteriorment, es mostrarà un seguit d'explicacions en un format. En aquest moment, se't donarà un formulari que hauràs d'omplir per analitzar la comprensió de les explicacions rebudes. Un cop omplert, es retirarà la explicació i el primer formulari i se't entregarà un segon formulari que hauràs de respondre amb total sinceritat."

Un cop exposada aquesta presentació i després de resoldre possibles dubtes del participant, començarà la prova d'avaluació. Llavors, s'exposarà una de les 6 situacions de sol·licituds de crèdit. Aquest exemple a continuació s'ha dut a terme amb el usuari número 1, exposat en el annex de contrafactuals.

"Ets una noia de 23 anys, que actualment es troba vivint a un poble a les afores de Barcelona. Per poder-te moure cap al teu lloc de treball a Barcelona, utilitzaves el teu cotxe, però aquest ha patit una averia que no permet arreglar-lo. Donat que no comptes amb un altre vehicle, decideixes demanar un crèdit al banc per adquirir un cotxe de segona mà. A la sol·licitud introdueixes, a part de l'edat i el sexe, la següent informació: indiques que sol·licites un crèdit de 900 euros a retornar en 18 mesos, indiques a més que el teu treball es requereix qualificació per realitzar-la, que disposes de un habitatge al teu nom, que disposes d'un petit compte d'estalvis i un compte corrent amb nivell adquisitiu baix i que el motiu de la sol·licitud de crèdit és adquirir un cotxe. Malauradament, el crèdit és classificat com a denegat i s'ofereixen les següents explicacions"

Un cop finalitzada la presentació, primer de tot, es preguntarà al participant si li ha sorgit algun dubte. Posteriorment, es procedeix a ensenyar en un dels dos formats les explicacions contrafactuals generades per a l'usuari, en aquest cas, els contrafactuals generats per a l'usuari 1.

6.4. Consentiment informat per a participar a la prova d'avaluació

DECLARACIÓ DE CONSTENTIMENT INFORMAT I PROCESSAMENT DE DADES PERSONAL

En aquest document, l'autor del treball Gerard March Moy informa sobre la informació recol·lectada amb motiu del desenvolupament del treball final de grau d'enginyeria informàtica **Models d'intel·ligència artificial explicables i contrafactuals**.

El treball en qüestió es desenvolupa al voltant de la intel·ligència artificial i la capacitat de certs models de generar una explicació del resultat que retorna. Per al desenvolupament pràctic s'ha creat un sistema que retorna una predicció respecte els risc de oferir un crèdit a un client. Per posar a prova aquest sistema i avaluar les explicacions retornades, es farà ús d'aquestes dades per introduir-les al sistema i que aquest generi primer de tot una predicció del risc de concedir el crèdit i posteriorment generar una explicació del resultat obtingut.

Posteriorment, es realitzarà un procés d'avaluació, en el que es realitzarà una gravació de l'usuari per analitzar l'enteniment de l'explicació generada per part de l'usuari. Aquesta gravació servirà per analitzar el comportament de l'usuari i en el moment que s'hagi analitzat, **la vídeo gravació serà eliminada.**

Les dades que es sol·licitaran seran les següents:

- Edat
- Sexe
- Treball del sol·licitant
- Tipus d'habitatge (habitatge comprat, llogat o no compte amb habitatge)
- Comptes d'estalvi a nom del client.
- Nivell del compte corrent a nom del client.
- Quantitat de crèdit sol·licitat
- Temps de retorn del crèdit.
- Motiu de la sol·licitud del crèdit

Les dades que es recolliran per al desenvolupament del projecte s'emmagatzemaran de manera anònima fins la presentació del treball final de grau i només serà visualitzada per el realitzador del treball, el tutor responsable del treball i posteriorment el tribunal, que avaluarà el resultat final del projecte, entre juny i juliol de 2024.

A partir d'aquest punt, es mencionaran els drets que tindrà el participant en tot moment:

- La participació en aquest projecte serà **voluntari** i **no serà remunerada**.
- El participant en qualsevol moment podrà notificar la seva decisió de no continuar amb la seva participació en el projecte i podrà retirar-se sense cap conseqüència.
- En cas de qualsevol dubte, pots contactar al desenvolupador principal del projecte Gerard March Moy a través del correu gmarchmo7@alumnes.ub.edu.

DECLARACIÓ DE CONSTENTIMENT INFORMAT I PROCESSAMENT DE DADES PERSONAL

Jo, _____, amb DNI _____ - en endavant, el/la usuari(a) – dono el meu consentiment per participar en el projecte **Models d'intel·ligència artificial explicables i contrafactuals** – en endavant, el projecte – que forma part de un estudi que realitza Gerard March Moy – en endavant, l'avaluador – per al treball final de grau del grau d'Enginyeria Informàtica, impartit per la Universitat de Barcelona (UB) amb objectiu d'avaluar les explicacions obtingudes de un model d'intel·ligència artificial.

La informació recol·lectada 'emmagatzemarà de manera anònima i les dades no es podran relacionar amb la meva persona. Només aquelles persones involucrades en el desenvolupament del projecte podran tenir accés a les dades personals: l'avaluador, la responsable del treball final de grau i el tribunal responsable d'avaluar el treball.

Estic informat/-ada de que **no hi haurà cap retribució** per la participació en aquest estudi. També soc coneixedor de la possibilitat de negar-me a participar en el estudi i inclús de retirar-me en el moment que ho consideri oportú.

Per tan, declaro que:

☐ He llegit la informació relativa a aquest estudi

☐ Entenc que la participació és voluntària i puc abandonar l'estudi en qualsevol moment

☐ Dono el meu consentiment per a ser gravat durant la prova d'usabilitat. Estic informat de que aquesta gravació serà **eliminada** un cop s'hagi analitzat les gravacions.

Finalment, de forma voluntària, expresso que (escull una de les dues opcions):

☐ Dono el meu consentiment per a participar en el estudi proposat

☐ No dono el meu consentiment per a participa en el estudi proposat

Signat per _____, el ____ / ____ / _____:

6.5. Formulari de comprensió

Formulari de comprensió

Indica quin format s'ha mostrat durant la prova d'avaluació.

☐ Format tabular ☐ Format textual

Primera part: Indica per a cada una de les següents afirmacions si estàs d'acord, si estàs en desacord o bé no ho saps.

1. La informació que es mostra indica les variacions que es poden fer per a que es concedeixi el crèdit.
    a. Estic d'acord
    b. No estic d'acord
    c. No se
2. El número màxim de característiques de la sol·licitud que s'han de canviar no és mai superior a dos.
    a. Estic d'acord
    b. No estic d'acord
    c. No se

3. El motiu de la sol·licitud del crèdit pot canviar l'estat de la sol·licitud.
    a. Estic d'acord
    b. No estic d'acord
    c. No se
4. Tots els contrafactuals indiquen que la quantitat de crèdit ha d'augmentar per a que l'estat de la sol·licitud canviï.
    a. Estic d'acord
    b. No estic d'acord
    c. No se

5. Com a mínim es planteja un possible canvi en la duració del crèdit.
    a. Estic d'acord
    b. No estic d'acord
    c. No se

6. El lloc del treball pot causar que l'estat del crèdit canviï.
    a. Estic d'acord
    b. No estic d'acord
    c. No se

Segona part: respon a la següent pregunta de forma breu i concisa.
Pregunta: quins possibles canvis es proposa al client per a que el crèdit canviï el seu estat de denegat a concedit?

6.6. Qüestionari de satisfacció

Qüestionari de satisfacció

Indica quin format s'ha mostrat durant la prova d'avaluació.

☐ Format tabular ☐ Format textual

Indica amb un valor numèric entre 1 i 6 quant d'acord estàs amb les següents afirmacions, (1 = "molt en desacord", 6 = "molt d'acord"):

1. Les explicacions rebudes m'han resultat fàcils de comprendre

1	2	3	4	5	6

2. Les explicacions rebudes han resultat satisfactòries.

1	2	3	4	5	6

3. Les explicacions rebudes són suficientment detallades.

1	2	3	4	5	6

4. Les explicacions rebudes no compten amb informació no rellevant.

1	2	3	4	5	6

5. Les explicacions rebudes són molt complertes.

1	2	3	4	5	6

6. Les explicacions rebudes són útils per a entendre el motiu de la denegació del crèdit.

1	2	3	4	5	6

7. Les explicacions rebudes per obtenir el crèdit són precises

1	2	3	4	5	6

8. Les explicacions rebudes per obtenir el crèdit són fiables

1	2	3	4	5	6

7. Referències

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *Towards Robust Interpretability with Self-Explaining Neural Networks*. https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

Anderson, R. (2016). The Rashomon Effect and Communication. *Canadian Journal of Communication*, *41*(2), 249–270. https://doi.org/10.22230/cjc.2016v41n2a3068

Aussó, S., Dominguez, D., & Quintana, M. (2022). *Guia sobre l'explicabilitat en la Intel·ligència Artificial*. Funcació TIC Salut Social. https://ticsalutsocial.cat/wp-content/uploads/2023/01/230226-Informe-Explicabilitat.pdf

Berber, A., & Srećković, S. (2023). When something goes wrong: Who is responsible for errors in ML decision-making? *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01640-1

Bove, C., Lesot, M.-J., Tijus, C. A., & Detyniecki, M. (2023). Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 188–203. https://doi.org/10.1145/3581641.3584082

Christoph Molnar. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Independently published. https://christophm.github.io/interpretable-ml-book/

CoworkingSpain. (2024). https://weareCloudworks.com/coworking-barcelona/

Decision Tree. (2024). https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., & Das, P. (2018). *Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives*. http://arxiv.org/abs/1802.07623

Gnat, S. (2021). Impact of Categorical Variables Encoding on Property Mass Valuation. *Procedia Computer Science*, *192*, 3542–3550. https://doi.org/10.1016/j.procs.2021.09.127

Hale, K. (2021). A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied. *Forbes*. https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/

Hans Hofmann. (1994, November 16). *UCI Machine Learning Repository*. Statlog (German Credit Data). https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*. http://arxiv.org/abs/1812.04608

IBM. (2019). *AI Explainability 360 Documentation*. https://aix360.readthedocs.io/en/latest/?badge=latest#

IBM. (2023). *AIX360* (0.3.0). GitHub. https://github.com/Trusted-AI/AIX360

IBM research trusted AI. (n.d.). *AI Explainability 360*. https://aix360.res.ibm.com/

Kiwi Remoto. (2024, May 28). Suelo de Desarrollador/a Inteligencia Artificial. https://www.kiwiremoto.com/sueldo/desarrollador-inteligencia-artificial/

Lombrozo, T. (2012). Explanation and Abductive Inference. In *The Oxford Handbook of Thinking and Reasoning* (pp. 260–276). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Luss, R., Chen, P.-Y., Dhurandhar, A., Sattigeri, P., Zhang, Y., Shanmugam, K., & Tu, C.-C. (2019). *Leveraging Latent Features for Local Explanations*. http://arxiv.org/abs/1905.12698

March, G. (2024). *TFGAvaluacioDeContrafactuals*. https://github.com/Gerard01mm/TFGAvaluacioDeContrafactuals

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Elsevier*. https://doi.org/10.1016/j.artint.2018.07.007

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558

Pawelczyk, M., Bielawski, S., van den Heuvel, J., Richter, T., & Kasneci, G. (2021a). *CARLA - Counterfactual And Recourse Library*. https://github.com/carla-recourse/CARLA

Pawelczyk, M., Bielawski, S., van den Heuvel, J., Richter, T., & Kasneci, G. (2021b). *Carla's Documentation*. https://carla-counterfactual-and-recourse-library.readthedocs.io/en/latest/

Real Academia Española. (2023). Explicación. In *Diccionario de la lengua española* (23.7). real Academia Española. https://dle.rae.es/explicaci%C3%B3n

Real Academia española. (2023). Interpretar. In *Diccionario de la lengua española* (23.7). Real Academia Española. https://dle.rae.es/interpretar

Simkus, J. (2023, July 31). *Simply Psychology*. Between-Subjects Design: Overview & Examples. https://www.simplypsychology.org/between-subjects-design.html

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

sklearn.tree.DecisionTreeClassifier. (2024). https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

The Linux Foundation. (2024). *The Linux Foundation AI & Data*. https://lfaidata.foundation/

Violet Turri. (2022, January 17). *Carnegie Mellon University, Software Engineering Institute's Insights (blog)*. What Is Explainable AI? https://insights.sei.cmu.edu/blog/what-is-explainable-ai/

Wang, X., & Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650