Generació de dades sintètiques per a l'entrenament de models d'IA de predicció del risc cardiovascular

JIALIANG YE YAN

***Resumen*—Amb l'increment de les tecnologies amb intel·ligència artificial (IA), és necessari entrenar-les amb dades. Però aquestes dades són limitades, pel fet que una IA entrenada amb més dades tindrà una major efectivitat en la seva resolució. El repte principal és obtenir un conjunt de dades (*dataset*) gran per a poder entrenar una IA. Per tal d'assolir-ho, és necessari crear dades sintètiques per a l'entrenament d'aquestes tecnologies. Durant el desenvolupament de l'article es donen detalls de la generació de dades sintètiques en Python que s'utilitzaran per a entrenar una IA que detecta el risc cardiovascular, a més a més de l'aplicació creada per aquest entrenament i els seus potencials beneficis.**

***Index Terms*—Dataset, Scikit-learn, Python, FastAPI, Flask.**

I. INTRODUCCIÓ

En l'àmbit de la informàtica, està agafant molt de pes l'ús de les tecnologies amb intel·ligència artificial (IA). Un dels àmbits, entre molts, que està implementant la intel·ligència artificial és a la medicina, ja que és una eina amb molt de potencial per a tots els professionals sanitaris en les anàlisis dels pacients per a les possibles malalties .

Per això, el present projecte analitzarà i proposarà solucions per a la generació de dades sintètiques per a l'entrenament de models d'IA de predicció de risc cardiovascular.

Les dades sintètiques són dades generades que simulen les característiques de les dades reals. En el camp de la medicina, són utilitzats per:

- **Completar conjunts de dades incomplets:** omplir espais en blanc dels conjunts de dades existents, obtenint una conjunts de dades més complet i útil per a la seva anàlisis.
- **Protegir la privacitat dels pacients:** crear conjunts de dades amb l'objectiu d'investigar i desenvolupar la ciència sense haver de comprometre la confidencialitat de les dades dels pacients.
- **Diversitat en el conjunt de les dades:** crear conjunts de dades més diversos, amb l'objectiu de reduir els biaixos dels algoritmes d'aprenentatge automàtic.

Degut a les grans quantitats de dades que es tenen guardats de tots el pacients, es pot considerar que en la medicina disposa de Big Data. El Big Data es refereix al gran conjunt de dades que pot arribar a ser complicada la seva gestió, processament o l'anàlisi d'aquestes mitjançant eines convencionals. [1]
De la mateixa forma, si s'analitza tot el conjunt de dades, es pot arriba a extreure informació molt valuosa per a l'entitat implicada. Alguns dels avantatges que et pot proporcionar l'anàlisi del conjunt de dades són:

- **Reducció de costos:** a partir de la informació obtinguda de l'anàlisi de les dades es pot identificar formes de reduir els costos.
- **Millor eficiència:** permet identificar àrees on es pot millorar els procediments.
- **Presa de decisions:** al tenir una millor comprensió de les dades, permet realitza millor preses de decisió.
- **Creació de nous serveis:** a partir de la informació obtinguda de les dades es pot crear serveis per satisfer les necessitats dels clients.

En l'àmbit de la medicina, el Big Data és utilitzat pels següents casos:

- **Millorar la comprensió de les malalties:** identificant els factors de risc, els patrons de progressió i els possibles tractaments de les malalties.
- **Desenvolupar nous medicaments o teràpies:** personalitzant els tractaments a cada una dels pacients de forma individual.
- **Millorar l'atenció primària:** millorar l'eficiència dels sistemes i reduint els costos.

En el projecte s'utilitzarà el llenguatge de programació Python, perquè és una de les millors solucions per treballar sobre intel·ligència artificial, ja que aquest tipus de llenguatge va ser creat per a l'anàlisi de dades. [2]

Per a la creació de les dades sintètiques, es basa en models obtinguts de Kaggle o de SyntheticMAss. [3] [4]

L'algoritme que s'utilitzarà en l'entrenament de la intel·ligència artificial en el projecte serà de l'anàlisi dels diferents models que ofereix la llibreria de codi obert Scikit-learn. [5]

I per finalitzar el projecte, es realitza un aplicació web per a que els usuaris puguin realitzar les seves prediccions. Per això es fan servir els frameworks de Flask [6] i FastAPI [7].

I-A. Tipologia y paraules clau

Dataset, Sobre-ajustament.

I-B. Definicions, acrònims i abreviacions

- **Dataset:** Conjunt de dades, estructurades o no estructurades, utilitzades per a l'entrenament d'un algoritme d'intel·ligència artificial.
- **Sobre-ajustament:** característica d'un model d'intel·ligència artificial en el que el model entrenat s'ha ajustat massa a les dades entrenades. Provocant així

que les prediccions amb dades noves, resultin errònies. Per tant el model és incapaç de generalitzar les dades d'entrada i predir de forma efectiva.

II. QUE ÉS LA IA?

Intel·ligència artificial és un dels conceptes més buscats en el mercat d'avui en dia. Però què és la IA?

Segons John McCarthy, "*és la ciència i l'enginyeria per a crear màquines intel·ligents, especialment programes informàtics intel·ligents. Està relacionada amb la tasca similar d'utilitzar ordinadors per a comprendre la intel·ligència humana, però la IA no té per què limitar-se a mètodes que siguin biològicament observables*". [8]

Un altre definició de John McCarty sobre la intel·ligència artificial és "*fer que una màquina es comporti d'una manera que seria considerada intel·ligent en un ésser humà*".

També, els informàtics Andreas Kaplan i Michael Haenlein, defineixen la intel·ligència artificial com "*la capacitat que té un sistema per interpretar dades externes correctament, aprendre d'aquestes dades, i fer servir els coneixements adquirits per completar tasques i assolir objectius específics mitjançant una adaptació flexible*". [9]

Per tant, tenint en compte les definicions anteriors, es pot extreure la conclusió que la intel·ligència artificial es basa en la configuració d'una màquina que rep un volum gran de dades inicials per assolir un comportament final desitjat. En el cas del desenvolupament de l'article, és rebre un gran volum de dades inicials ja validades, realitzar un entrenament amb les dades i generar perfils sintètics que coincideixin amb les dades creades per tal de continuar amb l'entrenament.

II-A. Orígens

Des dels inicis fins a l'actualitat la IA ha passat per diferents etapes. El primer és l'origen de la IA, que és el test de Turing. El test de Turing, és una prova proposada pel matemàtic Alan Turing en el 1950 en el que es basa a discernir si una màquina té un comportament intel·ligent.

Arran d'aquest fet, en el 1956, John McCarthy, Marvin Minsky y altres informàtics de l'època, van proposar les bases de la intel·ligència artificial com a disciplina en la Conferència de Darthmouth.

Durant les dècades del 50 al 60, es van produir diversos avenços com el primer robot a substituir a un humà en una línia d'acoblament o el primer xatbot a poder establir una conversa amb un humà.

A aquests fets el precedeix una època en el que no hi ha grans avenços, degut al poc finançament dels projectes basats en intel·ligència artificial. Però la comunitat no va estar parat i es van enfocar a desenvolupar sistemes experts, que es basa a rebre coneixement humà per després prendre decisions basades en la informació rebuda.

El punt d'inflexió per reprendre la IA i fins a l'auge d'avui en dia va ser l'aprenentatge automàtic (*Machine Learning*) en els 90. A més a més, es disposava de moltes més dades per a poder realitzar un millor entrenament. Algunes de les fites més importants van ser la victòria d'una màquina al campió d'escacs Garry Kasparov en el 1997, l'assistent virtual d'Apple Siri en el 2011, la victòria d'una màquina al campió del món d'AlphaGo Ke Jie en el 2017 o el reconegut ChatGPT en el 2022. [10] [11]

II-B. Estat de l'art

Avui en dia, la intel·ligència artificial ha realitzat molts avenços. Una d'elles és la intel·ligència artificial generativa, que poden arribar a crear cada cop imatges, vídeos, àudio o text més "perfectes". Entenent perfecte com a similar a un realitzar per un humà. La IA generativa es pot trobar en tecnologies actuals com els xatbots, assistents virtuals, educadors o creadors dels continguts esmentats anteriorment.

Un altre dels avenços és la IA multimodal, que consisteix a combinar diferents tipus de dades per obtenir una experiència. Tecnologies que apliquin aquesta IA són els cotxes autònoms, sistemes de realitat virtual o eines d'anàlisi de dades.

També s'utilitza la IA en la ciberseguretat per a prevenir amenaces cibernètiques com la detecció d'intrusions, anàlisi de programari maliciós i la protecció de les dades mitjançant la identificació i el xifratge de les dades més sensibles.

I en altres sectors, com en el financer amb la detecció de fraus o la presa de decisions; o en el sector sanitari amb la detecció de malalties mitjançant IA, on les dades sintètiques estan prenent protagonisme.

III. ENTRENAMENT DE LA IA

L'entrenament d'una IA es basa en 5 passos.

- **Obtenció de les dades:** definir el problema a resoldre i l'obtenció de les dades necessàries per obtenir els resultats desitjats.
- **Preprocessament de les dades:** correcció de dades errònies, dades irrellevants, normalització de dades i separació de la quantitat de dades per entrenar el model i la quantitat de dades per provar el model.
- **Selecció del model:** elecció del model d'intel·ligència artificial a utilitzar. Hi ha molts models diferents.
- **Entrenament del model:** adaptació dels paràmetres del model per obtenir el millor resultat possible.
- **Avaluació del model:** dur a terme proves de validació i mesurar el rendiment.

Correu de contacte: jialiangye96@gmail.com

Document tutoritzat per: Remo Suppi Boldrito

III-A. Format d'emmagatzematge de les dades

Per al preprocessament i l'entrenament del model d'IA, s'ha de tenir clar el tipus de dades a utilitzar. Entre les diferents possibilitats d'emmagatzematge de les dades, tenim:

- Dades estructurades.
- Dades no estructurades.

Les dades estructurades són aquelles que dades, que es guarden amb un format predefinit, normalment solen ser emmagatzemats en format text. Destaquem les següents estructures d'emmagatzematge:

- **CSV (Comma-Separated Values):** fitxer altament compatible, que es guarda en format text pla i amb una estructura de fila-columna on els camps se separen per coma, punt i coma o tabulació.
- **XML (Extensible Markup Language):** fitxer que utilitza etiquetes per generar l'estructura de les dades, les dades es guarden de forma estructurada, disposa d'una sintaxi formada i documentada i els documents amb format xml poden ser processats per molts programes.
- **JSON (JavaScript Object Notation):** les dades es representen en format clau-valor dins de '' i separats per coma, les dades que guarda pot ser molt variat (text, número, booleans, null, arrays o objectes), utilitza una estructura anidada, els arrays en JSON es representen amb '[]', és fàcil de llegir i molt utilitzat en aplicacions web.
- **BSON (Binary JSON):** format que guarda les dades en un format binari, l'estructura és la mateixa que la del JSON, però amb la diferència que JSON les guarda en format text pla.

I les dades no estructurades, són aquelles en què les dades no són emmagatzemades amb cap format predefinit. Les estructures de dades poder ser les següents:

- **Text sense format:** contingut pot ser variat i complex en el que el significat depèn del context en el qual es troba i són complexes de processar.
- **Imatge:** gran varietat del format de les imatges on cada una té les seves característiques i complexitat alta en analitzar i extreure informació de milers o de milions de píxels de la imatge.
- **Àudio**: gran varietat de guardat dels àudios on cada una té les seves peculiaritats i la informació es guarda en un format temporal. Com les anteriors dades no estructurades, són complexes d'analitzar.
- **Vídeo**: disposa d'una gran varietat de guardat on cada una té les seves característiques i són molt complexes d'analitzar. A diferència dels formats mencionats, els vídeos guarda la informació en un format tant espacial com temporal.

Entre altres, ja que també existeixen dades no estructurades per a àmbits específics com les dades de les xarxes socials, dades geoespacials, dades genètiques, logs de màquines, etc.

En el nostre cas les dades generades de forma sintètica, s'emmagatzemaran en el format estructurat **CSV**. Ja que és un format altament compatible amb les eines que s'utilitzen avui en dia.

IV. GENERACIÓ DE DADES SINTÈTIQUES

La generació sintètica de les dades en l'àmbit de la medicina pren cada cop més rellevància, ja amb l'avanç de les tecnologies, et permet realitzar entrenaments més complets d'algoritmes per ajudar als sanitaris en l'anàlisi i tractament de malalties. [12] [13] Les principals raons del perquè les dades sintètiques han pres importància, són:

1. **Privacitat:** la informació mèdica és molt sensible i el seu ús ha de regir-se per normes de confidencialitat. Les dades sintètiques permeten crear conjunts de dades que simulen la informació dels pacients sense haver de revelar les dades personals identificables. Això és crucial per a protegir la privacitat dels pacients i complir amb les regulacions de privacitat de les dades com el RGPD (Reglament General de Protecció de Dades).
2. **Investigació i desenvolupament:** la recerca mèdica és sovint obstaculitzada per la falta de dades o per la dificultat d'obtenir consentiments dels pacients per utilitzar les seves dades. Les dades sintètiques poden suplir aquesta manca, permetent als investigadors la creació de grans conjunts de dades per a entrenar models d'aprenentatge automàtic, provar algoritmes i desenvolupar noves tecnologies mèdiques, sense posar en risc la privacitat dels pacients.
3. **Millorar la precisió dels models d'IA:** els models d'IA necessiten grans quantitats de dades d'alta qualitat per a entrenar-se. Les dades sintètiques poden ajudar a augmentar i diversificar els conjunts de dades d'entrenament, la qual cosa pot resultar en models més precisos i robustos.
4. **Escenaris hipotètics:** les dades sintètiques es poden utilitzar per a la creació d'escenaris hipotètics que serien difícils o impossibles de replicar en el món real. Això permet als investigadors explorar diferents possibilitats i provar noves intervencions en un entorn segur i controlat.
5. **Reduir costos i temps de desenvolupament:** la recopilació de dades mèdiques reals pot ser un procés costós i lent. Les dades sintètiques poden oferir una alternativa més ràpida i econòmica, permetent desenvolupar i implementar noves tecnologies mèdiques de manera més eficient.

En general, les dades sintètiques tenen el potencial de revolucionar la medicina permetent una recerca i un desenvolupament més ràpids, innovadors i ètics. No obstant això, és important destacar que les dades sintètiques presenten alguns desafiaments, com la necessitat de garantir la seva qualitat i representativitat, així com la possibilitat de biaixos algorítmics. A mesura que la tecnologia continua desenvolupant-se, serà crucial abordar aquests desafiaments per garantir que les dades sintètiques s'utilitzin de manera responsable i beneficiosa per a la salut pública. [14]

Per tant, la generació sintètica de les dades per a l'entrenament del nostre model d'IA, ha d'estar limitat a uns paràmetres reals, per tal d'obtenir un dataset el més similar a la realitat possible, obtenint així una IA que determina amb

Correu de contacte: jialiangye96@gmail.com

Document tutoritzat per: Remo Suppi Boldrito

més precisió el risc cardiovascular. Els paràmetres utilitzats en la creació dels datasets són:

- *age:* indica l'edat del pacient.
- *gender:* indica el sexe del pacient.
- *ethnic:* indica L'ètnia a la que pertany el pacient.
- *poverty* indica si el pacient pateix un estat de pobresa.
- *smoke:* indica si el pacient és fumador.
- *ascvd:* indica si el pacient pateix de la malaltia cardiovascular ateroscleròtica.
- *icm:* indica l'index d massa corporal del pacient.
- *mellitus_diabetis:* indica si el pacient pateix de diabetis mellitus.
- *hypertension:* indica si el pacient pateix d'hipertensió.
- *dyslipidemia:* indica si el pacient pateix de dislipemia.
- *familial_hypercholesterolemia:* indica si el pacient pateix de hipercolesterolemia familiar.
- *alco:* indica si el pacient és alcohòlic.
- *atrial_fibrillation:* indica si el pacient pateix de fibrilación auricular.
- *anxiety_disorder:* indica si el pacient pateix d'un trastorn d'ansietat.
- *depressive_disorder:* indica si el pacient pateix d'un trastorn depressiu.
- *psychosis:* indica si el pacient pateix de psicosis.
- *colt:* indica el nivell de colesterol total, la suma del colesterol de baixa densitat amb el colesterol d'alta densitat.
- *ldl:* indica el nivell de colesterol de baixa densitat.
- *tg:* indica la quantitat de triglicèrids del pacient.
- *antidepressants:* indica si el pacient pren ansiolítics.
- *antidiabetic_treatment:* indica si el pacient està en un tratament antidiabètic.
- *antihypertensive_treatment:* indica si el pacient està en un tractament antihipertensiu.
- *lipid_lowering_treatment:* indica si un pacient està en un tractament hipolipemiant.
- *chronic_renal_failure:* indica si el pacient pateix d'insuficiència renal crònica.
- *renal_replacement_therapy:* indica si el pacient esta en un tractament per un trasplantament renal.
- *kidney_transplant:* indica si el pacient ha rebut un trasplantament renal.
- *COVID_history:* indica si el pacient ha patit el COVID.
- *anemima:* indica si el pacient pateix d'anèmia.
- *chronic_obstructive:* indica si el pacient pateix d'una malaltia obstructiva crònica.
- *severe_obstructive_sleep_apnea:* indica si el pacient pateix d'apnea.
- *fatty_liver:* indica si el pacient té un fetge gras.
- *erectile_dysfunction:* indica si el pacient pateix d'una disfunció erèctil.
- *rheumatoid_arthritis:* indica si el pacient pateix d'artritis reumatoide.
- *migraines:* indica si el pacient pateix de migranyes.
- *systemic_lupus_erythematosus:* indica si e pacient pateix de lupus eritematós sistèmic.
- *alzheimer:* indica si el pacient pateix d'alzheimer.
- *systolic_blood_pressure:* indica la pressió sanguina sistòlica del pacient.
- *score:* indica la probabilitat del risc cardiovascular.

En el projecte no es profunditza en els paràmetres, ja que són terminologies de l'àmbit sanitari. Si vol obtenir més informació, consulta la Real Academia Nacional de Medicina de España. [15]

IV-A. Restriccions del dataset

Per tal d'obtenir el dataset encara més similar al món real, és necessari configurar les caracterítiques dels camps de dades, per tal que les proporcions d'aquestes siguin el més semblant possible a les del món real. Que tal com s'ha esmentat en el punt anterior, és per obtenir una IA millor entrenada i obtenir resultats més encertats. Les restriccions utilitzades per a la creació del dataset estan disponibles en l'annex d'aquest document.

IV-B. Generació del dataset

Per a la generació del dataset, s'utilitza el llenguatge de programació Python, que juntament amb les llibreries que ofereix el llenguatge poden generar de forma sintètica els valors per als perfils dels pacients. Les llibreries de Python que s'utilitzen són:

- **csv:** per a poder guardar les dades en el format escollit, que és el CSV.
- **random:** per poder generar un perfil de forma aleatòria i que no segueixi un patró en concret.
- **math:** utilitzada per al càlcul del valor que determina el risc cardiovascular.
- **scipy.stats:** en concret s'utilitza la funció de skewnorm, que genera valors amb una distribució normal.

Per a la generació satisfactòria del perfil d'un pacient, amb totes les dependències, s'ha de passar fins a 6 iteracions on cada una d'elles s'encarrega de generar els valors adequats i de forma aleatòria.

- El primer pas és la creació de tots els camps del dataset i assignant valors a aquells que són independents.
- En la segona iteració es generen els valors dels camps que, depenent del gènere del pacient, presenten un percentatge o un altre de patir la malaltia.
- En la tercera iteració de la generació del perfil, es generen els valors que depenen de l'edat del pacient.
- En la següent iteració es generen els valors que depenen de si el pacient pateix de insuficiència renal crònica.
- El penúltim pas és l'assignació del valor del camp en el que indica si el pacient ha rebut un trasplantament renal.
- I l'últim pas abans de tenir el perfil generat, és calcular el valor del camp score, ja que és el que determina el risc cardiovascular.

Per acabar, es guarden els valors obtinguts a un document en format CSV.

Correu de contacte: jialiangye96@gmail.com

Document tutoritzat per: Remo Suppi Boldrito

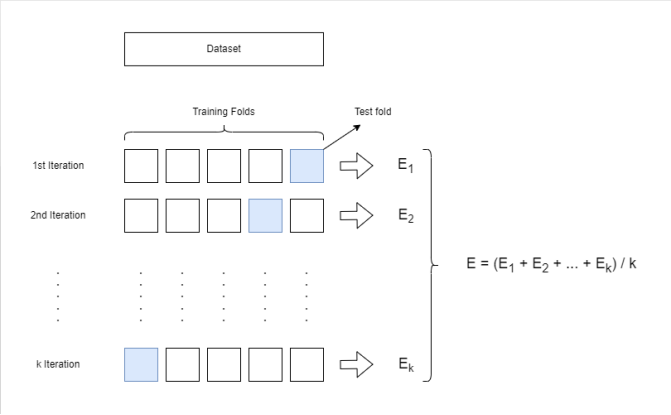

Figura 1: Mètode K-Fold Cross Validation

V. ALGORITME D'INTEL·LIGÈNCIA ARTIFICIAL

Per l'elecció de l'algoritme a utilitzar en l'entrenament de la IA, fem servir la llibreria Scikit-learn, que és una llibreria del llenguatge de programació Python de codi obert.

V-A. Dataset

A partir del dataset sobre malalties cardiovasculars, obtinguda de Kaggle [3], el passem per un procés de validacio de les dades mitjançant el KFold.

El Kfold és una tècnica utilitzada en l'aprenentatge automàtic per a avaluar el rendiment d'un model. És una validació creuada, que és una estratègia per a estimar el rendiment d'un model en les prediccions sobre un conjunt de dades diferent del qual es va utilitzar per a entrenar el model.

Aquesta tècnica consisteix a separar el dataset en k subconjunts, anomenats folds. Després, el model s'entrena k vegades, usant un fold diferent com a conjunt de proves cada vegada. El rendiment del model correspon a la mitjana del càlcul de l'entrenament de cada un dels folds, d'aquesta forma s'obté una estimació més robusta del rendiment del model, tal i com es pot observar en la Figura 1.

Llavors el passem per algun dels algoritmes supervisats que disposa la llibreria Scikit-learn, que són: [5]

- *Linear Models*: es basa en determinar un algoritme en el que assumeix una relació lineal entre las variables de entrada (independents) y el valor de sortida (variable dependent).
- *Linear and Quadrantic Discriminant Analysis*: algoritme que utilitza dades basades en una distribució normal amb una matriu de covariància.
- *Kernel ridge regression*: combina la regresió de cresta (ridge regression) amb el mètode de kernel per determinar relacions no lineals entre les variables. On la regressió de cresta és una variant d'un model lineal en el que hi ha una regularització de les dades per evitar el sobre-ajustament del resultat final. I el mètode kernel et permet manipular resultats sense haver de realitzar transformacions de dades perquè els resultats no són separables de forma lineal o és necessari transformar les dades a un espai de major dimensió.
- *Support Vector Machines*: algoritme que determina un hiperpla per separa les dades en un espai de característiques, maximitzant l'espai entre les característiques. D'aquesta forma es pot generalitzar, classificant així dades noves, mai vistes anteriorment.
- *Stochastic Gradient Descent*: calcula el gradient a partir d'una petita part de les dades. El gradient és un indicador de com i cap a on canvien els paràmetres del model, millorant el rendiment d'aquest.
- *Nearest Neighbors*: predicció d'un punt de dades a partir dels punts de dades més pròxims dins l'espai de característiques.
- *Gaussian Processes*: enfocament probabilístic, estimant la incertesa de les prediccions.
- *Naive Bayes*: algoritme basat en el teorema de Bayes, per predir la classe a partir de les característiques observades en les dades durant l'entrenament.
- *Decision Trees*: utilitzen una estructura d'arbre per prendre decisions, basat en es característiques de les dades proporcionades. En l'estructura d'arbre, els nodes representen la característica i el vèrtex d'un node a un altre representa una decisió basada en aquella característica.
- *Ensembles*: basat en combinar diversos models per tal d'obtenir un millor rendiment en les prediccions. Alguns exemples són: *Gradient boosting, random forests, bagging, voting, stacking*. En concret ens fixem amb *random forests*, que es basa en crear molts arbres, on cada arbre s'entrena d'una forma diferent. Posteriorment es combinen els resultats, creant així un resultat que redueix el sobre-ajustament.
- *Multiclass and multioutput algorithms*: algoritme utilitzat en models en el que el resultat pot tenir més de dos clases o més d'una variable resultant.
- *Semi-supervised learning*: es basa en l'entrenament a partir de dades etiquetades i dades no etiquetades.
- *Isotonic regression*: algoritme utilitzar per determinar tendències ascendents o descendents.
- *Neural network models (supervised)*: basat en imitar el funcionament del cervell humà, composat per capes de neurones interconnectades.

En concret, es trien el model *Nearest Neighbors*, el *Random Forest* que pertany a *Ensembles* i el model *Neural network models (supervised)* per compara l'eficàcia d'aquestes.

Per comparar els resultats obtinguts de realitzar la cross validation, comparem els resultats amb els resultats que s'obtenen d'una llibreria Python anomenada Pycaret, que agilitza el procés de tria del model a entrenar, realitzant una comparació amb tots els models existents d'IA en el mercat.

V-B. Resultats

Per a poder comparar els resultats d'executar la validació Kfold als diferetns models, ens fixarem a les següents mètriques:

- **MSE:** de sigles en anglès de Mean Squared Error. És una mètrica utilitzada per a avaluar el rendiment dels models de regressió. Mesura la diferència quadràtica mitja dels valors predits amb els valors reals en un conjunt de dades. Per tant, indica que tan lluny estan les prediccions del seu model dels valors reals. Un valor de MSE més petit, on el mínim és 0, indica un millor ajustament del model i, per tant, significa que les prediccions del model estan més prop dels valors reals. Per contra, un valor de MSE més alt, no té un valor límit, indica un ajustament pobre, cosa que significa que les prediccions del model estan més allunyades dels valors reals.
- **MAE:** de les sigles en anglès Mean Absolute Error. És una mètrica utilitzada per a avaluar el rendiment dels models de regressió. Mesura la diferència quadràtica mitja dels valors predits amb els valors reals en un conjunt de dades. Per tant, indica que tan lluny estan les prediccions del seu model dels valors reals, però en termes de magnitud absoluta. Un valor més petit, pròxim a '0', indica un millor ajustament, cosa que significa que les prediccions del model estan més prop dels valors reals en termes de magnitud absoluta. Per contra, un valor més gran indica un ajustament més pobre, cosa que significa que les prediccions del model estan més allunyades dels valors reals en termes de magnitud absoluta.
- **RMSE:** de les sigles en anglès Root Mean Squared Error. És una mètrica, com en els casos anteriors, utilitzada per a avaluar el rendiment dels models de regressió. És l'arrel quadrara del valor obtingut en la MSE. Com més petita el valor de la RMSE, més ajustada estarà el model per realitzar prediccions més properes a les reals.
- **R-squared (R2):** també conegut com a coeficient de determinació, és una mètrica utilitzada per a avaluar el rendiment d'un model de regressió lineal. Mesura la variància entre les variables dependents (variable a predir) respecte a la variable o variables independents en el model. En altres paraules, indica què tan bé el model explica les variacions en les dades observades. Com més pròxima, el valor de la mètrica, estigui del '1', millor realitzarà les prediccions.

Després de comprovar l'efectivitat dels models escollits, podem veure els resultats en la següent Figura 2.

Segons les comprovacions realitzades, s'obtenen resultats poc representatius, ja que tots els valors de les mètriques estan lluny de ser valors acceptables. Ja que, perquè MSE, MAE, RMSE siguin valors bons, aquests han de ser molt pròxims a 0 i el valor de la mètrica R-squared ha de ser pròxim a 1.
Comparem els resultats obtinguts amb els resultats que ens proporciona l'eina de Pycaret, en la següent Figura 3.

Tal i com es pot comprobar amb l'eina Pycaret, ens proporciona resultats una mica millors però amb models que la llibreria de Scikit-learn no té a la seva disposició.

Per tant caldrà realitzar més estudis per determinar el motiu dels valors tan llunyans dels desitjats.

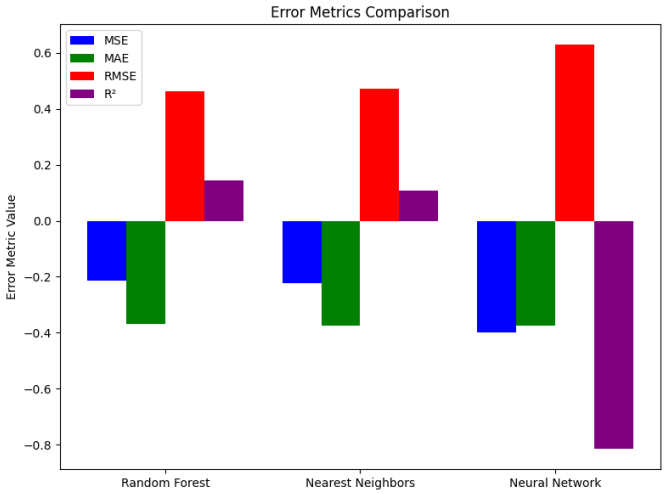

Figura 2: Resultats K-Fold Cross Validation

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.3643	0.1812	0.4257	0.2748	0.2987	0.3675	1.7910
lightgbm	Light Gradient Boosting Machine	0.3616	0.1815	0.4260	0.2738	0.2988	0.3647	0.2930
catboost	CatBoost Regressor	0.3607	0.1821	0.4268	0.2712	0.2992	0.3638	5.2970
ada	AdaBoost Regressor	0.3739	0.1880	0.4336	0.2477	0.3040	0.3791	0.1790
lr	Linear Regression	0.3996	0.1965	0.4433	0.2137	0.3093	0.4060	0.8210
ridge	Ridge Regression	0.3996	0.1965	0.4433	0.2137	0.3093	0.4060	0.0340
lar	Least Angle Regression	0.3996	0.1965	0.4433	0.2137	0.3093	0.4060	0.0360
br	Bayesian Ridge	0.3997	0.1965	0.4433	0.2137	0.3093	0.4061	0.0360
rf	Random Forest Regressor	0.3687	0.1991	0.4462	0.2033	0.3127	0.3708	7.4800
en	Elastic Net	0.4198	0.2036	0.4512	0.1853	0.3157	0.4236	0.0340
lasso	Lasso Regression	0.4295	0.2069	0.4549	0.1721	0.3192	0.4328	0.0400
llar	Lasso Least Angle Regression	0.4295	0.2069	0.4549	0.1721	0.3192	0.4328	0.0290
et	Extra Trees Regressor	0.3694	0.2132	0.4618	0.1468	0.3229	0.3716	4.8570
knn	K Neighbors Regressor	0.3749	0.2230	0.4722	0.1078	0.3297	0.3857	0.2070
huber	Huber Regressor	0.4124	0.2310	0.4797	0.0759	0.3260	0.4234	0.4530
omp	Orthogonal Matching Pursuit	0.4724	0.2362	0.4860	0.0548	0.3416	0.4766	0.0310
dummy	Dummy Regressor	0.5000	0.2500	0.5000	-0.0002	0.3516	0.5044	0.0300
dt	Decision Tree Regressor	0.3699	0.3699	0.6082	-0.4799	0.4216	0.3741	0.1700
par	Passive Aggressive Regressor	0.4847	0.3928	0.6254	-0.5718	0.4225	0.6532	0.0860

Figura 3: Resultats de la comparació del models d'IA amb Pycaret

VI. APLICACIÓ WEB

Per tal de que els profesionals sanitaris puguin interactuar amb l'eina, es crearà una aplicació web seguint l'arquitectura REST API, tal i com es pot observar en la Figura 4.

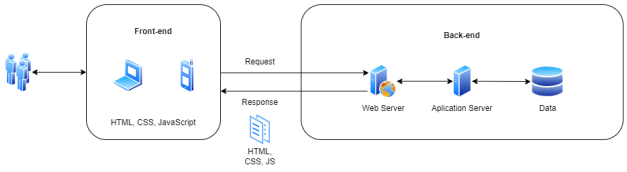

Figura 4: Arquitectura aplicació web

L'aplicació web estarà codificada a partir de Python, amb el framework Flask [6] per realitzar el Front-end i utilitzant el framework de FastAPI [7] per realitzar la part del Back-end.

VI-A. Motivació

Els motius per la tria dels frameworks de Flask i FastAPI, són que, Flask et proporciona una tecnologia molt consolidada en el mercat i que té a la seva disposició una gran comunitat d'informàtics. El motiu per la tria de Flask per realitzar el front-end, són:

- **Simplicitat i facilitat d'ús:** arquitectura minimalista, corba d'aprenentatge suau i una sintaxi intuïtiva que facilita la creació d'aplicacions web bàsiques.
- **Flexibilitat i extensibilitat:** és altament flexible i s'adapta a diversos projectes, permet la integració amb diferents biblioteques i eines de JavaScript populars com React, Vue.js i Angular, facilitant la creació d'interfícies dinàmiques i interactives i suporta plantilles HTML per a crear vistes atractives i personalitzades.
- **Entorn familiar per a desenvolupadors de Python:** ofereix un entorn familiar per als desenvolupadors Python, facilitat d'aquesta forma la transició al desenvolupament web.
- **Gran comunitat i suport:** Flask compta amb una comunitat àmplia i activa de desenvolupadors que brinden suporti contribueixen al desenvolupament del framework.
- **Accesible:** és un framework lleuger i amb una corba d'aprenentatge suau que permet als desenvolupadors principiants crear prototips funcionals i aprendre els fonaments del desenvolupament web sense necessitat de dominar conceptes complexos.

Tots aquests punts han conclòs amb l'elecció de Flask per realitzar el front-end de l'aplicació web.
I el motiu per triar FastAPI per realitzar el back-end,són:

- **Alt rendiment i escalabilitat:** construït utilitzant tecnologies d'alt rendiment, la qual cosa li permet manejar grans volums de sol·licituds de manera eficient i és ideal per a aplicacions web que demanden un alt rendiment i escalabilitat, com APIs en temps real o serveis web amb molts usuaris.
- **Productivitat i simplicitat:** sintaxi intuïtiva basada en anotacions i decoradors, la qual cosa facilita la creació de APIs i la connexió amb bases de dades.
- **Validació de dades integrada:** eines integrades per a validar i assegurar la integritat de les dades que s'envien i reben per la API. Això ajuda a prevenir errors i protegir les dades sensibles.
- **Documentació automàtica:** genera automàticament documentació interactiva per a la API, facilitant la seva comprensió i ús per part d'altres desenvolupadors i clients.
- **Flexibilitat:** funciona en diferents sistemes operatius i entorns d'execució.
- **Suport a diferents esquemes de dades:** admet diferents formats de dades populars com JSON, YAML i FormUrlEncoded.
- **Seguretat:** inclou característiques de seguretat integrades com a protecció contra atacs CSRF i validació de signatures JWT.
- **Fàcil de realitzar proves:** facilita l'escriptura de proves unitàries i d'integració per a garantir la qualitat i confiabilitat de la API.

I és per aquests motius que s'ha triat FastAPI per realitzar el back-end.

VI-B. Front-end

El front-end de l'aplicació, que és el contingut que veurà el client per el dispositiu electrònic (telèfon mòvil, tauleta, ordinador portàtil, ordinador sobretaula, etc.). Això està format per HTML, CSS i JavaScript.

- **HTML:** fitxer amb un llenguatge que defineix l'estructura i el contingut de la pàgina web.
- **CSS:** fitxer que conté llenguatge que determina l'estil visual de la pàgina web.
- **JavaScript:** llenguatge de programació que afegeix interactivitat al contingut estàtic de la pàgina web com càrrega de dades o les animacions.

VI-C. Back-end

El back-end és la part de l'aplicació web que s'encarrega de la lògica i del processament de les peticions dels clients per generar les respostes del contingut demanat. El back-end està format per les següents parts:

- **Infraestructura:** compost per servidors webs, sistemes operatius, gestors de bases de dades i eines de seguretat.
- **Bases de dades:** sistemes de gestió de dades per a emmagatzemar dades de l'aplicació i permet al back-end l'accés d'aquestes.
- **Llenguatges de programació:** llenguatge utilitzar per a realitzar tota la lògica de l'aplicació i per a processar les peticions rebudes per la part del client.
- **Frameworks:** eina que proporciona una estructura que simplifica el desenvolupament d'aplicacions web back-end.
- **Serveis web:** mitjançant API RESTful per a que les aplicacions o sistemes es puguin comunicar.

VI-D. Resultats

VII. PLANIFICACIÓ

El projecte dura un total de 22 setmanes, on la primera setmana comença el 19 de febrer de 2024 i l'última setmana correspon a la setmana del 15 de juliol de 2024, que seria la setmana 22 del projecte, tal i com es pot visualitzar en el Diagrama de Gantt que es troba en l'Annex.

VII-A. Fases del projecte

El projecte es pot segmentar el les següents fases, on cada una d'elles es realitza alguna de les tasques previstes.

- **Inici del TFG:** anàlisi previ del treball a realitzar durant el projecte. Entre les qual destaca el curs de Python per familiaritzar-me amb el llenguatge de programació, buscar informació de la llibreria Scikit-learn i també recordar les funcionalitats de les llibreries Pandas i Numpy.
- **Generació de perfils sintètics:** creació de perfils sintètics en Python amb els paràmetres establerts i l'emmagatzematge de les dades creades en el format acordat.
- **Entrenament de la IA:** elecció de l'algoritme a utilitzar i realitzar l'entrenament per a la detecció dels risc cardiovascular.
- **Aplicació web:** creació d'una aplicació web amb un front-end i un back-end per a l'entrenament de l'algoritme, introducció de dades i obtenció de resultats.
- **Proposta informe final:** entrega de l'informe final redactat amb el format demanat, per a que el tutor del projecte el pugui revisar i realitzar algunes correccions si són necessaries.
- **Proposta presentació:** entrega de la presentació a realitzar durant la defensa del projecte, per si cal realitzar alguna modificació.
- **Lliurament dossier:** entrega de tots els documents relacionats amb la realització del projecte
- **Lliurament pòster:** entrega del disseny del poster del projecte.
- **Defensa TFG:** exposició oral del projecte davant del tribunal per corroborar la realització del projecte.

VIII. References

Referencias

[1] "TEMA 1. BIG DATA EN SALUD. HACIA UNA SALUD PREDICTIVA Y PERSONALIZADA — salusplay.com." https://www.salusplay.com/apuntes/apuntes-de-salud-digital/tema-1-big-data-en-salud-hacia-una-salud-predictiva-y-personalizada.

[2] "wingsoft.com." https://www.wingsoft.com/blog/mejores-lenguajes-IA.

[3] "Cardiovascular Disease dataset — kaggle.com." https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

[4] J. N. A. Q. C. M. D. H. C. D. K. D. T. G. S. M. Jason Walonoski, Mark Kramer, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, SyntheticMass." https://synthea.mitre.org/downloads.

[5] "User guide: contents — scikit-learn.org." https://scikit-learn.org/stable/user_guide.html.

[6] "Welcome to Flask &x2014; Flask Documentation (3.0.x) — flask.palletsprojects.com." https://flask.palletsprojects.com/en/3.0.x/.

[7] "FastAPI — fastapi.tiangolo.com." https://fastapi.tiangolo.com/.

[8] "www-formal.stanford.edu." https://www-formal.stanford.edu/jmc/whatisai.pdf.

[9] "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence — sciencedirect.com." https://www.sciencedirect.com/science/article/abs/pii/S0007681318301393?via%3Dihub.

[10] N. Rodríguez, "La Historia de la Inteligencia Artificial: Desde sus Orígenes hasta el Presente — natisr." https://medium.com/@natisr/historia-de-la-inteligencia-artificial-63277f78fe2c.

[11] "Artificial Intelligence Timeline Infographic – From Eliza to Tay and beyond — digitalwellbeing.org." https://digitalwellbeing.org/artificial-intelligence-timeline-infographic-from-eliza-to-tay-and-bey

[12] "Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies — ncbi.nlm.nih.gov." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059086/.

[13] "Synthetic data in health care: A narrative review — ncbi.nlm.nih.gov." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931305/.

[14] "Synthetic data in medical research — ncbi.nlm.nih.gov." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9951365/.

[15] "Real Academia Nacional de Medicina: Presentaci&xF3;n Diccionario de t&xE9;rminos m&xE9;dicos — dtme.ranm.es." https://dtme.ranm.es/index.aspx.

ANNEX

Variable	Descripció	Variable depenent	Requisit
age	edat del pacient en dies; valor numèric	No	Rang de valors: [7665,29200]
gender	0 = dona, 1 = home	No	0 = 51.02 %, 1 = 48.98 %
ethnic	0 = caucàsic, 1 = africà, 2 = asiàtic	No	caucàsic = 96.7 %, africà = 2.7 %, asiàtic = 0.6 %
poverty	0 = no pobre, 1 = pobre	gender	If (gender == 0) then 0 = 79 %, 1 = 21 %. If (gender == 1) then 0 = 77.8 %, 1 = 22.2 %.
smoke	0 = no fumador, 1 = fumador	No	0 = 79 %, 1 = 21 %
ascvd	0 = no ascvd permatura, 1 = ascvd prematura	No	0 = 95.5 %, 1 = 4.5 %
icm	distribució normal, valor numèric amb 2 decimals	No	(icm ¡25) = 32 %, (25 ¡= icm ¡= 30) = 39 %, (icm ¿30) = 29 %
mellitus_diabetis	0 = no mellitus diabetis, 1 = mellitus diabetis	No	0 = 89 %, 1 = 11 %
hypertension	0 = no hipertensió, 1 = hipertensió	gender	If (gender == 0) then 0 = 47 %, 1 = 53 %. If (gender == 1) then 0 = 64 %, 1 = 36 %.
dyslipidemia	0 = no dislipèmia, 1 = dislipèmia	No	0 = 42 %, 1 = 58 %
familial_hypercholesterolemia	0 = no hipercolesterolemia familiar, 1 = hipercolesterolemia familiar	No	0 = 95.5 %, 1 = 0.5 %
alco	0 = no pren alcohol, 1 = pren alcohol	No	0 = 82 %, 1 = 18 %
atrial_fibrillation	0 = no fibril·lació auricular, 1 = fibril·lació auricular	age	If (age ¿40) then 0 = 95.6 %, 1 = 4.4 %. If (age ¡= 40) then 0 = 99.93 %, 1 = 0.07 %.
anxiety_disorder	0 = no trastorn d'ansietat, 1 = trastorn d'ansietat	gender	If (gender == 0) then 0 = 91.2 %, 1 = 8.8 %. If (gender == 1) then 0 = 95.5 %, 1 = 4.5 %.
depressive_disorder	0 = no trastorn depressiu, 1 = trastorn depressiu	gender	If (gender == 0) then 0 = 74.1 %, 1 = 5.9 %. If (gender == 1) then 0 = 97.7 %, 1 = 2.3 %.
psychosis	0 = no psicosi, 1 = psicosi	No	0 = 98.8 %, 1 = 1.2 %
colt	nivell de colesterol total; valor numèric	No	(colt ¡200) = 49.5 %, (colt ¿= 200) = 50.5 %
ldl	nivell de colesterol ldl; valor numèric	No	(ldl ¡130) = 55.1 %, (ldl ¿= 130) = 44.9 %
tg	nivell de triglicèrids; valor numèric	gender	If (gender == 0) then (tg ¡150) = 88.3 %, (tg ¿= 150) = 11.7 %. If (gender == 1) then (tg ¡150) = 76.8 %, (tg ¿= 150) = 23.2 %.
antidepressants	0 = no ansiolítics o antidepressius, 1 = ansiolítics o antidepressius	No	0 = 79.6 %, 1 = 20.4 %
antidiabetic_treatment	0 = no tractament antidiabètic, 1 = tractament antidiabètic	gender	If (gender == 0) then 0 = 31 %, 1 = 69 %. If (gender == 1) then 0 = 34 %, 1 = 66 %.
antihypertensive_treatment	0 = no tractament antihipertensiu, 1 = tractament antihipertensiu	gender	If (gender == 0) then 0 = 27 %, 1 = 73 %. If (gender == 1) then 0 = 32 %, 1 = 68 %.
lipid_lowering_treatment	0 = no tractament hipolipemiant, 1 = tractament hipilipemiant	gender	If (gender == 0) then 0 = 58 %, 1 = 42 %. If (gender == 1) then 0 = 61 %, 1 = 39 %.
chronic_renal_failure	0 = no insuficiència renal crònica, 1 = insuficiència renal crònica	gender	If (gender == 0) then 0 = 87.1 %, 1 = 12.9 %. If (gender == 1) then 0 = 85.6 %, 1 = 14.4 %.
renal_replacement_therapy	0 = no tractament renal substitutiu, 1 = tractament renal substitutiu	chronic_renal_failure	If (chronic-renal-failure == 0) then 0 = 100 %, 1 = 0 %. If (chronic-renal-failure == 1) then 0 = 99.87 %, 1 = 0.13 %.
kidney_transplant	0 = no trasplant renal, 1 = trasplant renal	renal_replacement_therapy	If (renal-replacement-therapy == 0) then 0 = 100 %, 1 = 0 %. If (renal-replacement-therapy == 1) then 0 = 99.99 %, 1 = 0.01 %.
COVID_history	0 = no historial COVID, 1 = historial COVID	No	0 = 93.7 %, 1 = 6.3 %
anemima	0 = no anemima, 1 = anemima	No	0 = 72 %, 1 = 28 %
chronic_obstructive	0 = no malaltia obstructiva crònica, 1 = malaltia obstructiva crònica	gender	If (gender == 0) then 0 = 96.1 %, 1 = 3.9 %. If (gender == 1) then 0 = 85.7 %, 1 = 14.3 %.
severe_obstructive_sleep_apnea	0 = no apnea obstructiva del somni greu, 1 = apnea obstructiva del somni greu	gender	If (gender == 0) then 0 = 98.1 %, 1 = 1.9 %. If (gender == 1) then 0 = 93.2 %, 1 = 6.8 %.
fatty_liver	0 = no fetge gras, 1 = fetge gras	No	0 = 75 %, 1 = 25 %
erectile_dysfunction	0 = no disfunció erèctil, 1 = disfunció erèctil	gender = men	If (gender == 0) then 0 = 100 %, 1 = 0 %. If (gender == 1) then 0 = 87.9 %, 1 = 12.1 %.
rheumatoid_arthritis	0 = no artritis reumatoide, 1 = artritis reumatoide	No	0 = 98.93 %, 1 = 1.07 %
migraines	0 = no migranyes, 1 = migranyes	gender	If (gender == 0) then 0 = 83.3 %, 1 = 16.7 %. If (gender == 1) then 0 = 94 %, 1 = 6 %.
systemic_lupus_erythematosus	0 = no lupus eritematós sistèmic, 1 = lupus eritematós sistèmic	No	0 = 99.91 %, 1 = 0.09 %
alzheimer	0 = no alzheimer, 1 = alzheimer	age	If (age ¿60) then 0 = 95 %, 1 = 5 %. If (age ¡= 60) then 0 = 100 %, 1 = 0 %.
systolic_blood_pressure	pressió arterial sistòlica en sang; valor numèric	gender	If (gender == 0) then (systolic-blood-pressure ¡140) = 62.9 %, (systolic-blood-pressure ¿= 140) = 37.1 %. If (gender == 1) then (systolic-blood-pressure ¡140) = 50.1 %, (systolic-blood-pressure ¿= 140) = 49.9 %.
score	indicador del risc cardiovascular; valor numèric	age, colt, smoke, systolic_blood_pressure	-

Tabla I: Restriccions del dataset

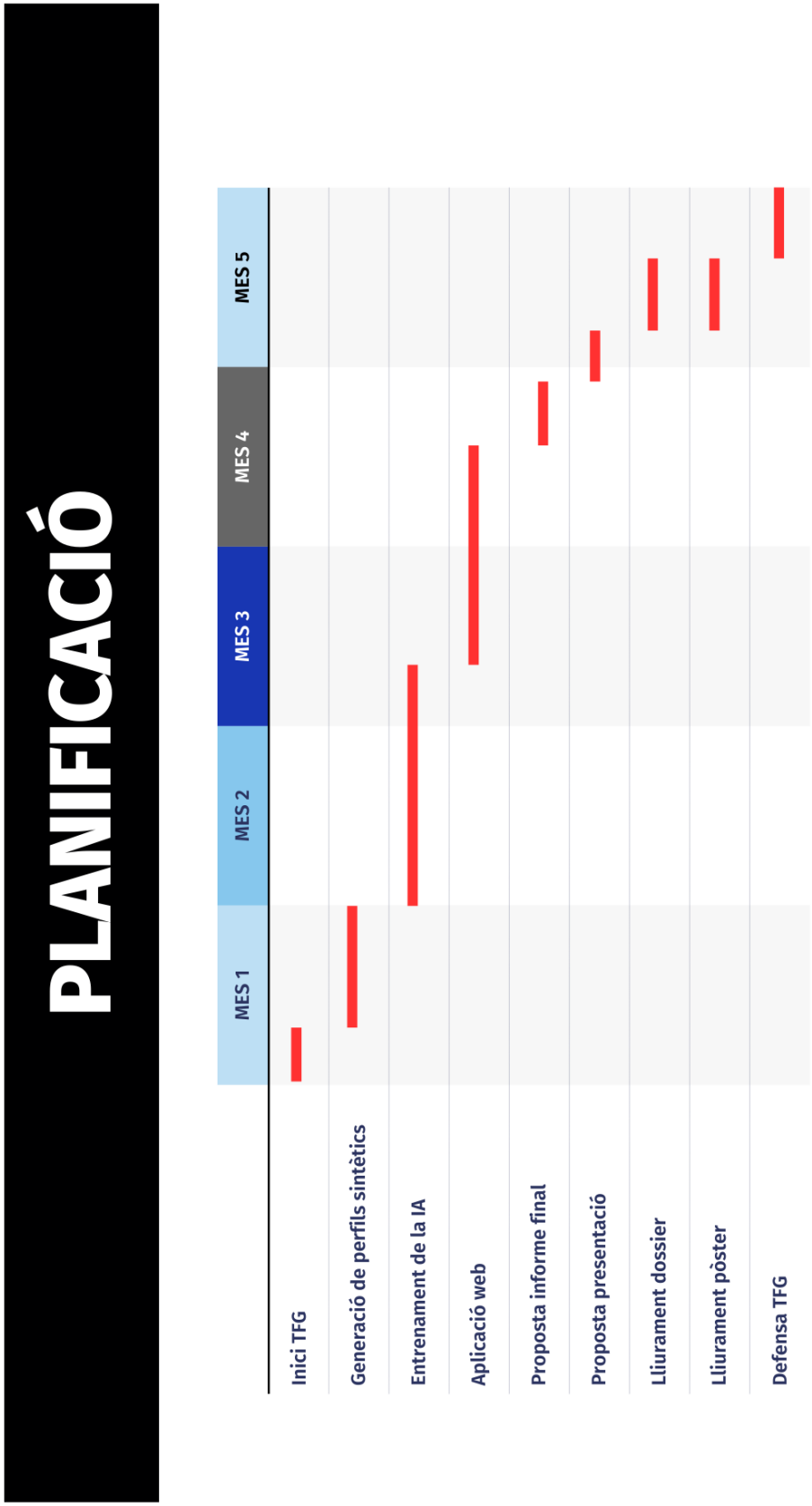

Figura 5: Planificació TFG