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. Una de les alternatives per assolir el repte són, la generació de 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.**

-Keywords: Dataset, Scikit-learn, Pycaret, IA, Dades Sintètiques.

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 la detecció 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]

En aquest projecte es faran servir eines i models que deriven d'altres datasets, en concret els obtinguts a Kaggle i SyntheticMAss. [3] [4] Per realitzar una anàlisis de predicción del risc cardiovascular bassat al dataset d'interès, es farà servir els diferents models implementats en PyCaret i SCIKitlearn. [5]

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*". [6]

Un altre definició de John McCarty sobre la intel·ligència artificial és "*fer que una màquina es comporti d'una manera*

Correu de contacte: jialiangye96@gmail.com
Document tutoritzat per: Remo Suppi Boldrito

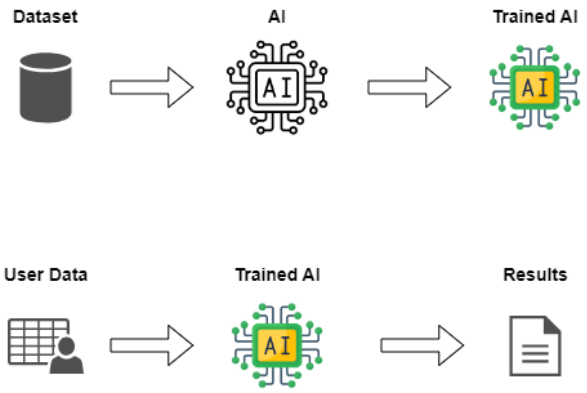

Figura 1: Funcionament bàsic de la IA

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*". [7]

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, tal i com es pot visualitzar en la Figura 1. 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. [8] [9]

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. 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. [10] [11] 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. [12]

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

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

En el projecte no es profunditza en els paràmetres, ja que són terminologies de l'àmbit sanitari.

III-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 1 d'aquest document.

De les restriccions del dataset, s'ha determinat que hi ha 5 tipus de creació del valor de la variable del perfil, que són:

- **Dos valors possibles:** aquells casos en el que la variable només presenta 2 possibles valors. El procés de generació del valor és:
  - Genera un valor random ente 0 i 1.
  - Compara el valor generat amb el percentatge per determinar el valor final.
- **Tres valors possibles:** aquells casos en el que la variable pot presentar fins a 3 possibles valors. El procés de generació del valor és:
  - Genera un valor random ente 0 i 1.
  - Compara el valor generat amb el rang del percentatge al que pertany per determinar el valor final.
- **Dependent del gènere del perfil:** aquells casos en el que la creació de la variable depèn del gènere del perfil. El procés de generació del valor és:
  - Genera un valor random ente 0 i 1.
  - Compara el valor generat amb el percentatge de la variable al que pertany al genere del perfil i determina el valor final a actualitzar al perfil creat.
- **Dependent de l'edat del perfil:** casos en el que la generació del valor ve determinat per l'edat del perfil. El procés de generació del valor és:
  - Genera un valor random ente 0 i 1.
  - Compara el valor generat amb el percentatge de la variable al que pertany a l'edat del perfil i determina el valor final a actualitzar al perfil creat.
- **Valor numèric:** casos en el que el valor que es crea és numèric i està dins un rang de valors determinat. El procés de generació del valor és:
  - Genera un valor random entre 0 i 1.
  - Compara el valor generat amb el rang del percentatge al que pertany i genera el resultat amb un numero enter random que pertany al rang de valors del percentatge al que pertany .

III-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.

III-C. Resultats

Per comprovar que el resultat obtingut, de la generació de les dades sintètiques, és vàlid amb els requisits proporcionats, que es troba a l'annex 2 del document, es realitza mitjançant una variable de tolerància.
Per tant, la metodologia implementada ha estat introduir la creació de les dades dins un bucle en el qual es realitzen els següents passos:

- **Generació de les dades sintètiques:** explicació del procediment en l'apartat anterior.
- **Procés de validació:** a partir de les proporcions i la variable de tolerància. Es valida variable per variable, a partir de la proporció desitjada i la tolerància, si la proporció dels valors es troben dins un rang acceptable. Si el valor es troba dins el rang, retorna un valor '0', del cas contrari retorna un '1'. El valor que retorna es guarda en un acumulador, que és el que indica si els valors són acceptables.
- **Comparació dels resultats:** si el resultat obtingut del procés de validació és que hi ha algun paràmetre que no compleix els requisits, es torna al pas inicial, que és la generació de les dades sintètiques i la seva posterior validació.
- **Guardat de les dades creades:** un cop validat les proporcions, es guarda el contingut creat en un document CSV, per al seu posterior ús.

Després de passar les dades generades per a la validació, obtenim els resultats que es pot visualitzar en l'annex 2.

IV. ALGORITME D'INTEL·LIGÈNCIA ARTIFICIAL

Per a predir el risc cardiovascular es buscarà el millor algoritme entrenat amb les dades generades i la llibreria Scikit-learn, que és una llibreria del llenguatge de programació Python de codi obert

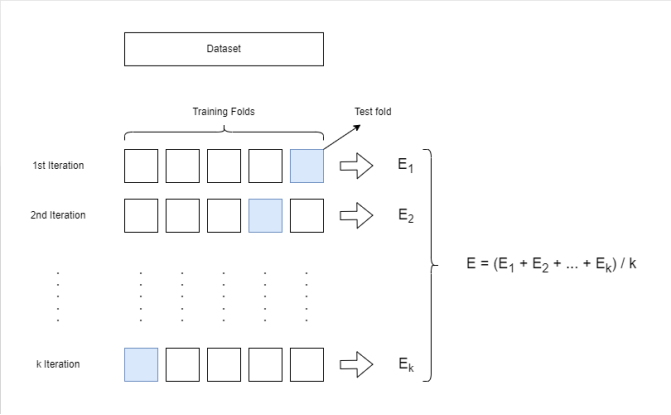

Figura 2: Mètode K-Fold Cross Validation

IV-A. Dataset

A partir del dataset sintètic, generat amb anterioritat, es crea una fórmula per tal d'aconseguir valors similars als que es troba en la taula del valor de SCORE2 d'Almirallmed [13], per poder provar la metodologia de l'entrenament del model de intel·ligència artificial.

La fórmula generada per calcular el valor d'Score2, s'obté a partir del valors de l'edat, el nivell del colesterol, la pressió arterial i de si el perfil és fumador.

Per a entrenar de forma correcta un model, es passen les dades sintètiques per un procés de validació 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 2.

L'anàlisi de quin és el millor algoritme supervisats es fa fent servir els que disposa la llibreria Scikit-learn [5]. De la gran quantitat d'algoritmes, s'ha escollit els 3 següents:

- *Nearest Neighbors*: predicció d'un punt de dades a partir dels punts de dades més pròxims dins l'espai de característiques.
- *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.
- *Neural network models (supervised)*: basat en imitar el funcionament del cervell humà, composat per capes de neurones interconnectades.

Per comparar els resultats obtinguts de realitzar la cross validation, els comparem 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.

IV-B. Resultats

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

- **MSE:** 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 pròxim a 0, indica un millor ajustament del model i, per tant, significa que les prediccions del model estan més prop dels valors reals.
- **MAE:** 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.
- **RMSE:** 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):** 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 Figura 3.

Segons les comprovacions realitzades, es pot deduir que el model de Random forest és la millor de les 3 opcions triades, perquè les mètriques de MAE, MSE, RMSE i R2 són les més properes als valors que indiquen un major encert. Mentres les mèstriques de MAE, MSE i RMSE, pel model Random Forest, són les més properes a 0 i la màtrica R2 és la més propera a 1 de les 3 opción.

Comparem els resultats obtinguts amb els resultats que ens proporciona l'eina de Pycaret, en la Figura 4.

Tal com es pot comprovar, 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.

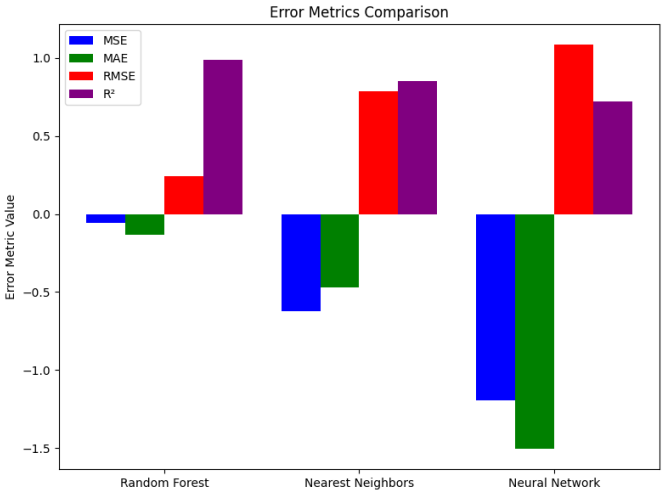

Figura 3: Resultats K-Fold Cross Validation

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.0368	0.0051	0.0694	0.9988	0.0096	0.0095	2.9730
lightgbm	Light Gradient Boosting Machine	0.0804	0.0235	0.1504	0.9946	0.0207	0.0199	0.2000
gbr	Gradient Boosting Regressor	0.1225	0.0457	0.2115	0.9894	0.0325	0.0326	0.5910
et	Extra Trees Regressor	0.1207	0.0551	0.2317	0.9872	0.0312	0.0287	2.2050
rf	Random Forest Regressor	0.1299	0.0613	0.2448	0.9858	0.0335	0.0307	2.4410
dt	Decision Tree Regressor	0.2351	0.1732	0.4152	0.9594	0.0605	0.0564	0.0650
knn	K Neighbors Regressor	0.5016	0.7007	0.8350	0.8367	0.1281	0.1237	0.0610
br	Bayesian Ridge	0.6482	0.8247	0.9061	0.8076	0.2194	0.2191	0.0360
ridge	Ridge Regression	0.6490	0.8252	0.9065	0.8075	0.2198	0.2195	0.0210
lr	Linear Regression	0.6490	0.8253	0.9065	0.8075	0.2198	0.2196	0.7400
lar	Least Angle Regression	0.6507	0.8271	0.9075	0.8070	0.2204	0.2205	0.0220
ada	AdaBoost Regressor	0.8271	0.8650	0.9297	0.7972	0.2135	0.3019	0.3580
en	Elastic Net	0.7233	1.0141	1.0051	0.7634	0.2269	0.2326	0.0290
lasso	Lasso Regression	0.7212	1.0164	1.0062	0.7629	0.2252	0.2309	0.0220
llar	Lasso Least Angle Regression	0.7212	1.0164	1.0062	0.7629	0.2252	0.2309	0.0200
omp	Orthogonal Matching Pursuit	0.7327	1.0451	1.0204	0.7561	0.2300	0.2358	0.0230
huber	Huber Regressor	0.6833	1.1941	1.0905	0.7217	0.1924	0.1915	0.2080
par	Passive Aggressive Regressor	0.7781	1.7554	1.3170	0.5924	0.2037	0.1899	0.0460
dummy	Dummy Regressor	1.5361	4.2788	2.0677	-0.0007	0.3908	0.4962	0.0310

Figura 4: Resultats de la comparació del models d'IA amb Pycaret

Per corroborar els valors obtinguts, comparem els resultats per a les diferents ponderacions de les variables de colesterol i de pressió arterial.

Primer comparem els resultats obtinguts amb un dataset en el qual el colesterol té poca importància en el càlcul final d'SCORE2 (Figura 5) i després se'l compara amb un dataset en el qual la pressió arterial té poca importància en el càlcul del valor d'SCORE2 (Figura 6).

De la mateixa forma, es passa els datasets per la llibreria de pycaret per obtenir valors i realitzar comprovacions. Primer pel dataset en què el colesterol té poca rellevància (Figura 7) i després per al que la pressió arterial té poca rellevància (Figura 8).

Amb totes les dades, podem determinar que el millor model d'IA que ofereix la llibreria Open Source Scikit-learn és el Random Forest. El codi d'aquest projecte es troba publicat al

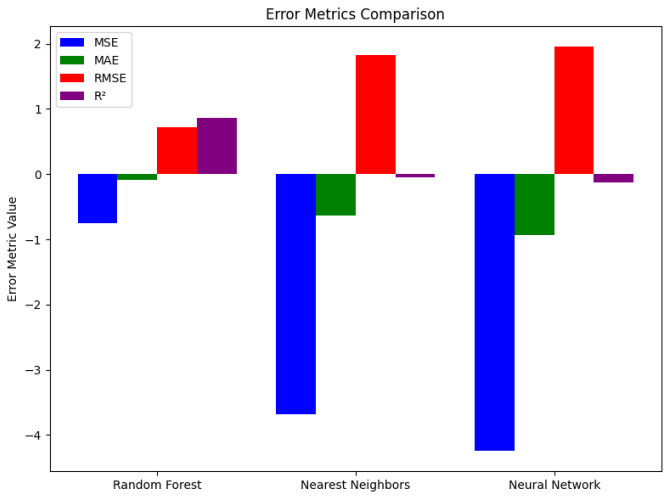

Figura 5: Resultats SCORE2 amb un baix valor de colesterol

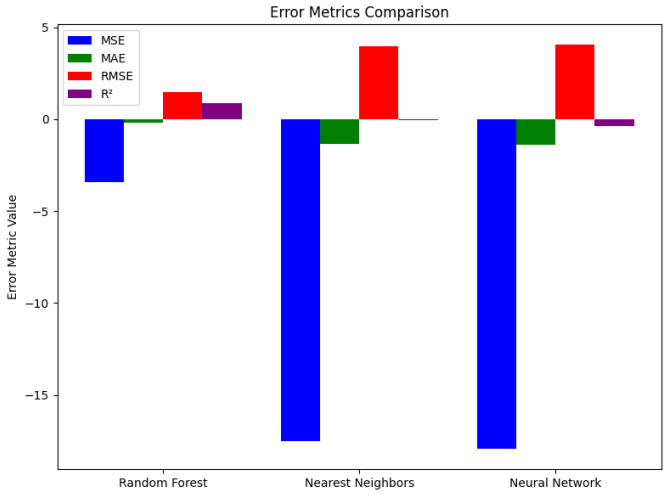

Figura 6: Resultats SCORE2 amb un baix valor de pressió arterial sitòlica

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.0605	0.4698	0.4680	0.9459	0.0327	0.0166	2.9930
et	Extra Trees Regressor	0.0900	0.8744	0.6560	0.8939	0.0497	0.0214	0.5490
rf	Random Forest Regressor	0.1171	1.1956	0.9162	0.7747	0.0709	0.0287	0.5930
lightgbm	Light Gradient Boosting Machine	0.1853	1.3991	0.9952	0.7435	0.1132	0.0836	0.1700
dt	Decision Tree Regressor	0.1625	1.3348	1.0506	0.6646	0.0919	0.0382	0.0270
gbr	Gradient Boosting Regressor	0.2402	2.2052	1.3319	0.4706	0.1362	0.1039	0.5640
br	Bayesian Ridge	0.6399	3.3706	1.6348	0.2779	0.2936	0.4384	0.0240
ridge	Ridge Regression	0.6458	3.3725	1.6363	0.2757	0.2961	0.4430	0.0200
lar	Least Angle Regression	0.6461	3.3726	1.6364	0.2756	0.2962	0.4432	0.0230
lr	Linear Regression	0.6461	3.3726	1.6364	0.2756	0.2962	0.4432	0.8850
lasso	Lasso Regression	0.7358	4.0281	1.8304	0.0570	0.3209	0.4832	0.0210
llar	Lasso Least Angle Regression	0.7358	4.0281	1.8304	0.0570	0.3209	0.4832	0.0220
en	Elastic Net	0.7454	4.0261	1.8307	0.0557	0.3245	0.4938	0.0200
omp	Orthogonal Matching Pursuit	0.7318	4.0714	1.8412	0.0454	0.3234	0.4768	0.0200
huber	Huber Regressor	0.5651	4.2148	1.8740	0.0118	0.3008	0.2766	0.1560
dummy	Dummy Regressor	0.6825	4.2604	1.8859	-0.0025	0.3140	0.4077	0.0210
par	Passive Aggressive Regressor	0.5370	4.3118	1.9020	-0.0257	0.3163	0.2277	0.0280

Figura 7: Resultats de Pycaret amb dataset amb un baix valor de colesterol

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.1450	2.8715	1.1249	0.9351	0.0532	0.0298	3.2460
et	Extra Trees Regressor	0.1962	4.6686	1.4480	0.8940	0.0617	0.0257	0.6460
rf	Random Forest Regressor	0.2502	5.8826	1.9606	0.7881	0.0904	0.0379	0.5910
lightgbm	Light Gradient Boosting Machine	0.3703	6.6036	2.0770	0.7680	0.1672	0.1470	0.1700
dt	Decision Tree Regressor	0.3098	5.8853	2.0373	0.7671	0.0999	0.0405	0.0260
gbr	Gradient Boosting Regressor	0.5287	10.3061	2.8598	0.4791	0.2023	0.1913	0.5630
br	Bayesian Ridge	1.3893	16.6174	3.5980	0.2650	0.4263	0.8738	0.0240
ridge	Ridge Regression	1.4027	16.6281	3.6014	0.2628	0.4277	0.8822	0.0200
lar	Least Angle Regression	1.4033	16.6286	3.6015	0.2627	0.4277	0.8826	0.0230
lr	Linear Regression	1.4033	16.6286	3.6015	0.2627	0.4277	0.8826	0.7380
en	Elastic Net	1.5183	19.1563	3.9429	0.0892	0.5074	0.9500	0.0190
lasso	Lasso Regression	1.5924	19.5925	4.0035	0.0545	0.5294	0.9893	0.0210
llar	Lasso Least Angle Regression	1.5924	19.5925	4.0035	0.0545	0.5294	0.9893	0.0210
omp	Orthogonal Matching Pursuit	1.5815	19.7051	4.0155	0.0489	0.5298	0.9758	0.0200
huber	Huber Regressor	1.0405	20.6845	4.1185	-0.0017	0.4380	0.3310	0.2030
dummy	Dummy Regressor	1.4520	20.6497	4.1183	-0.0019	0.4902	0.8007	0.0190
par	Passive Aggressive Regressor	1.1730	20.7355	4.1522	-0.0293	0.4925	0.4381	0.0240

Figura 8: Resultats de Pycaret amb dataset amb un baix valor de pressió arterial sistòlica

Github [14].

V. PLANIFICACIÓ

El projecte es va plantejar amb una durada total de 22 setmanes, on la primera setmana va ser la setmana del 19 de febrer de 2024 i l'última setmana va ser 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 3.

VI. CONCLUSIONS

La importància de les dades sintètiques en l'àmbit de la salut és alta, ja que sinó s'hauria de tractar amb les dades de pacients reals, motiu pel qual pot comprometre la integritat de la informació privada dels pacients afectats. Però amb les dades sintètiques, et permet generar dades fictícies que simulen un cas i realitzar estudis sobre aquesta, permetent així avançar més de pressa les investigacions de l'àmbit de la salut.

Per tant, durant el transcurs del projecte s'ha generat de forma satisfactòria un conjunt de dades sintètiques, que segueix les proporcions especificades en una fulla de requisits. Per a la validació de les dades sintètiques, es comprova les proporcions de les dades amb els requisits especificats. Ja que, en generar el dataset, l'assignació dels valors és aleatori i no sempre es pot complir el que està escrit en el codi. Per tant, amb la validació, es pot corroborar que les dades són les més semblants a les descrites en els requisits.

Amb el dataset sintètic validat, el passem per a alguns dels models d'IA que es troba a la llibreria Scikit-learn mitjançant una cross validation de les dades, en concret els models escollits ha estat el Random Forest, Nearest Neighbors i Neural Networks. Comparem els resultats obtinguts amb els resultats que ens ofereix la llibreria de Pycaret. D'aquesta forma, es corrobora els resultats obtinguts de realitzar un Kfolding al nostre dataset sintètic.

A partir de les mètriques obtingudes dels diferents models, es determina que el Random Forest és el millor. Ja que, dels tres models analitzats, és el que aporta millors resultats perquè és el que presenta un valor més baix pel que fa a errors en els valors predits respecte als valors reals i presenta el valor més alt pel que fa a precisió.

En definitiva, les dades sintètiques et permet avançat el desenvolupament d'un projecte en el que és necessari un dataset i no es vol posar en compromís les dades dels pacients o en el cas de que no es disposa d'un dataset molt gran.

Gràcies a la generació del dataset sintètic, s'ha pogut dur a terme la metodologia de l'entrenament d'un algormitme d'IA i poder analitzar els resultats en un cas en concret, que és la detecció del risc cardiovascular.

VII. LÍNIES FUTURES

Els següents passos a realitzar en un futur, poden ser:

- Afegir variables correlacionades al dataset sintètic, per aportar més similitud al món real.
- Ajustar la fórmula del càlcul d'SCORE2. Consultar professionals per obtenir la fórmula del càlcul del valor SCORE2 per obtenir un dataset més complet i així entrenar el model d'IA. Determinar les varaibles de dataset que influiex en el càlcul d'SCORE2 i determinar el valor exacte de les varaibles externes que s'utilitzen per al càlcul d'SCORE2. [15]
- Realitzar un entenament amb un dataset de dades real per comprovar resultats.
- Creació d'una aplicació web per a que els professionals sanitaris puguin realitzar les prediccions amb el model entrenat o la generació i obtenció del dataset sintètic.

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

[7] "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.
[8] 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.
[9] "Artificial Intelligence Timeline Infographic – From Eliza to Tay and beyond — digitalwellbeing.org." https://digitalwellbeing.org/artificial-intelligence-timeline-infographic-from-eliza-to-tay-and-beyond/.
[10] "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/.
[11] "Synthetic data in health care: A narrative review — ncbi.nlm.nih.gov." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931305/.
[12] "Synthetic data in medical research — ncbi.nlm.nih.gov." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9951365/.
[13] "Tema 2. El SCORE2 y las categorías de riesgo cardiovascular - Atención Primaria — atencionprimaria.almirallmed.es." https://atencionprimaria.almirallmed.es/tema-2-el-score2-y-las-categorias-de-riesgo-cardiovascular/.
[14] https://github.com/Jordiyy/tfg.
[15] "2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC) — academic.oup.com." https://academic.oup.com/eurheartj/article/42/34/3227/6358713?login=false#302948452.

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

variable	porcentaje_esperado (%)	resultado (%)		variable	porcentaje_esperado (%	resultado (%)
gender_women	51.02	51.3		men_tg_more_150_mg/dL	23.2	23.06
gender_men	48.98	48.7		antidepressants	20.4	20.11
caucassian_ethnic	96.7	96.53		antidiabetic_treatment	69	68.81
african_ethnic	2.7	2.98		antidiabetic_treatment	66	65.36
casian_ethnic	0.6	0.49		women_antihypertensive_treatment	73	73.1
women_poverty	21	21.05		men_antihypertensive_treatment	68	68.38
men_poverty	22.2	21.38		women_lipid_lowering_treatment	42	42.26
smoke	21	20.9		men_lipid_lowering_treatment	39	39.14
ascvd	4.5	4.69		women_chronic_renal_failure	12.9	13.02
icm_less_25	32	32.8		men_chronic_renal_failure	14.4	14.27
icm_between_25_30	39	38.07		renal_replacement_therapy	0.13	0.21
icm_more_30	29	29.13		kidney_transplant	0.01	0.01
mellitus_diabetis	11	29.13		covid	6.3	6.5
women_hypertension	53	52.46		anemima	28	28.47
men_hypertension	36	36.8		women_chronic_obstructive	3.9	4.27
dyslipidemia	58	57.62		men_chronic_obstructive	14.3	13.84
familial_hypercholesterolemia	0.5	0.38		women_severe_obstructive_sleep_apne	1.9	1.81
alco	18	18.01		men_severe_obstructive_sleep_apnea	6.8	6.9
atrial_fibrillation	0.7	0.59		fatty_liver	25	25.45
age_dependant_atrial_fibrillatic	4.4	4.63		women_erectile_dysfunction	0	0.0
women_anxiety_disorder	8.8	8.58		men_erectile_dysfunction	12.1	12.07
men_anxiety_disorder	4.5	4.5		rheumatoid_arthritis	1.07	1.0
women_depressive_disorder	5.9	6.28		women_migraines	16.7	17.02
men_depressive_disorder	2.3	2.69		men_migraines	6	5.85
men_depressive_disorder	1.2	1.15		systemic_lupus_erythematosus	0.09	0.05
colt_more_200_mg/dL	50.5	49.66		alzheimer_more_65	5	4.47
ldl_more_130_mg/dL	44.9	45.07		women_systolic_blood_pressure	37.1	37.04
women_tg_more_150_mg/dL	11.7	12.38		men_systolic_blood_pressure	49.9	49.55

Figura 9: Resultat de la validació de les variables

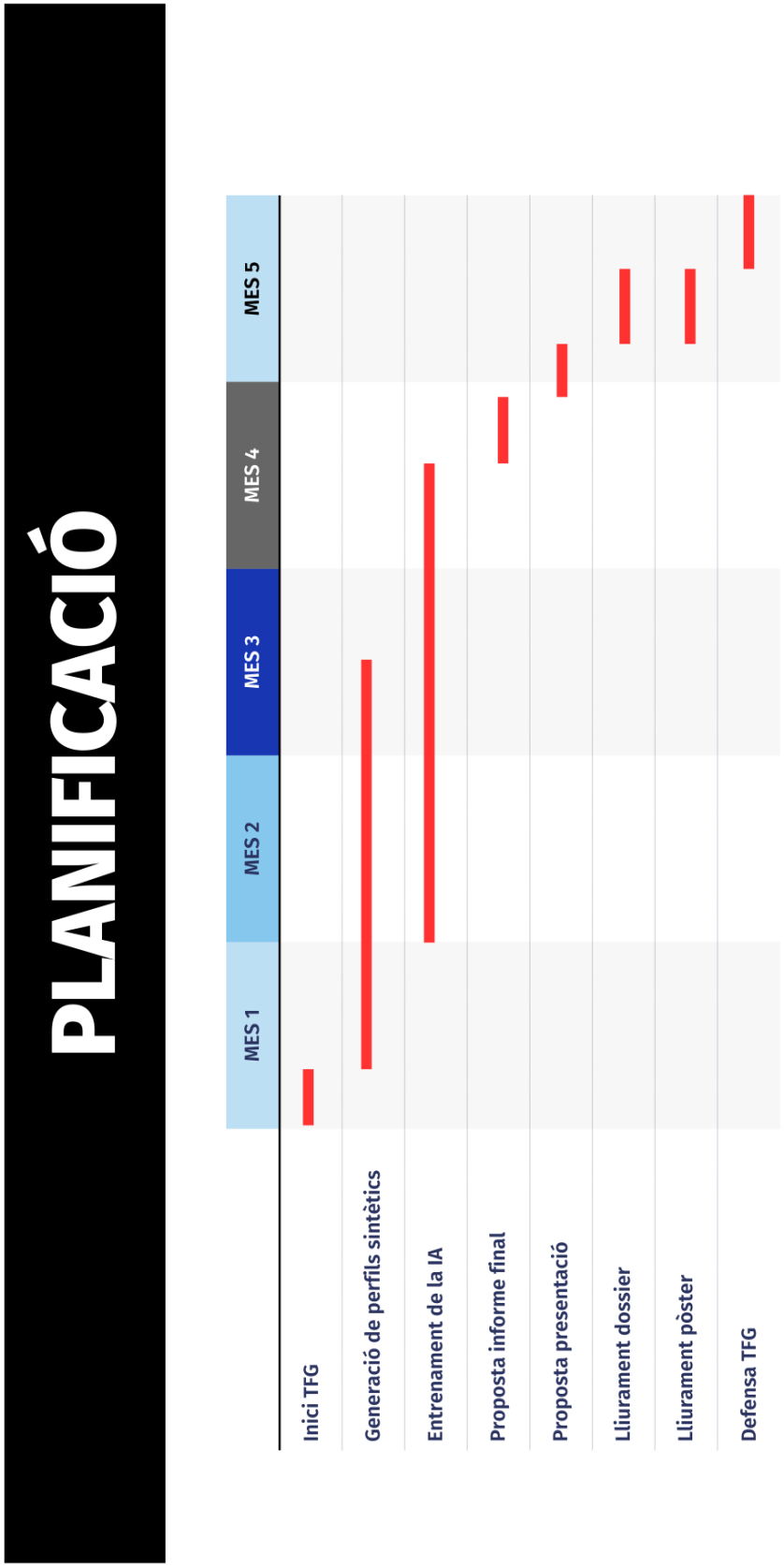

Figura 10: Planificació TFG