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

Jialiang Ye Yan *UAB*
Barcelona, España
1401633@uab.cat

***Resumen*—Amb l'increment de les tecnologies amb intel·ligència artificial (IA), és necessari entrenar-les amb dades. Però aquestes dades són limitades, i ja 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*—Paraules clau de l'article.**

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 pot aportar ajuda extra a tots els professionals sanitaris en les anàlisis dels pacients per a les possibles malalties .

Per això, l'article posarà èmfasi en la generació de dades sintètiques per a l'entrenament de models d'IA de predicció de risc cardiovascular.

Durant l'estudi previ a l'inici del projecte, una de les millors solucions per treballar sobre intel·ligència artificial és el llenguatge de programació Python, ja que aquest tipus de llenguatge va ser creat per a l'anàlisi de dades. [1]

Per a la generació sintètica de les dades, hem necessitat un conjunt de dades inicial sobre la qual treballar. Per tant, s'han definit el format del document d'emmagatzematge de les dades i les característiques que han de tenir d'aquestes, per tal de generar les dades sintètiques el més semblant a la realitat possible. [2] [3]

També es fa una comparació de tots els algoritmes que té a disposició la llibreria Scikit-learn per tal d'obtenir el millor algoritme a utilitzar. [4]

I acaba realitzant una aplicació amb una interfície gràfica, amb fastApi [5] y Flask [6], perquè els professionals mèdics puguin introduir les dades per obtenir el resultat, a més a més de permetre un entrenament continuat de l'algoritme escollit. [7] [8] [9]

I-A. Tipologia y paraules clau

Paraula 1, Paraula 2, Paraula 3, Paraula 4.

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.
- **Paraula n**: Definició paraula n.

II. QUE ÉS LA IA?

Intel·ligència artificial és un dels conceptes més sonats 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*". [10]

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*". [11]

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. [12] [13]

II-B. Estat de l'art

Avui en dia, la intel·ligència artificial ha realitzat molts avenços. Una d'elles es 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 en 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 sector sanitari amb la detecció de malalties mitjançant IA o la detecció de fraus o presa de decisions en el sector financer.

III. GENERACIÓ DE DADES SINTÈTIQUES

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.
- **Preprocesament de les dades**: correcció de dades errònies, dades irellevants, normalització de dades i separacio de la quantitat de dades per entrenar el model i la cuantitat de dades per provar el model.
- **Selecció del model**: elecció del model d'intel·ligència artificial a utilitzar. Hi han molts models diferents.
- **Entrenament del model**: adaptació dels paràmetes del model per obtenir el millor resultat possible.
- **Evaluació del model**: realitzar proves de validació i mesurar el rendiment.

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 es 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 processat per molts programes.
- **JSON (JavaScript Object Notation)**: les dades es representaen en format clau-valor dins de '' i separats per coma, les dades que guarda pot ser molt variat (text, numero, booleans, null, arrays o objectes), utilitza una estructura anidada, els arrays en JSON es representen amb '[]', és facil de llegir i molt utilitzat en aplicaions 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 de que JSON les guarda en format text pla.

I les dades no estructurades, són aquelles en que 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 complexe en el que el significat depèn del context en el que 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 vídoes 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.

III-B. Base de dades

Tria de base de dades per emmagatzemar el dataset.

- DB1: (caracterítiques de la DB)
- DB2: (caracterítiques de la DB)
- DB3: (caracterítiques de la DB)
- Etc.

III-B1. Caracterítiques del dataset:

La generació sintètica de les dades, ha d'estar limitat a uns paràmetres reals, per tal d'obtenir un dataset el més semblant a la realitat possible i per tant poder entrenar de forma més òptima la IA que determina el risc cardiovascular. Els paràmetres utilitzats en la creació dels datasets són:

- Parametre 1: (explicació).
- Parametre 2: (explicació).
- Parametre 3: (explicació).
- Parametre 4: (explicació).
- Parametre 5: (explicació).
- Etc.

III-C. Restriccions del dataset

Per tal d'obtenir el dataset encara més semblant al món real, és necessari implementar restriccions a les dades, per tal de que les proporcions de les dades siguin el més semblant al món real. Que tal i com s'ha esmentat en el punt anterior, és per obtenir una IA millor entrenada i obtenir resultats més encertat. Les restriccions són les següents: Taula I

IV. 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.

IV-A. Dataset

A partir d'un dataset sobre malalties cardiovasculars obtinguda a Kaggle, el separem en 2 parts. [2]

- El 80 % del dataset original serà utilitzar per a l'entrenament de l'algoritme.
- I el 20 % del dataset restant serà utilitzar per obtenir el percentatge d'encert.

Llavors el passem per algun dels algoritmes supervisats que disposa la llibreria Scikit-learn, que són: [4]

- *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.

Codi	Requisit	Prioritat
RES-001	(50 % homes, 50 % dones)	Essencial
RES-002	(altura mitja X amb desviació estàndard de Y)	Essencial.
RES-003	Descripció	Opcional.
RES-n	Descripció	Opcional.

Tabla I: Restriccions del dataset

En concret els passem pel *Logistic Regression* que pertany a *Linear Models*, el model *Nearest Neighbors*, *Decision Trees*, el *Random Forest* que pertany a *Ensembles* i el model *Neural network models (supervised)*.

IV-A1. Linear Models: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A2. Nearest Neighbors: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A3. Decision Trees: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A4. Ensembles: Gradient boosting, random forests, bagging, voting, stacking: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A5. Neural network models (supervised): L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-B. Resultats

Després de passar el mateix dataset d'entrenament a cada un dels algoritmes, obtenim que els que tenen un major percentajge d'encert són:

- Algoritme 1.
- Algoritme 2.
- Algoritme 3.

Tal i com es pot comparar en la següent gràfica d'encert.

V. APLICACIÓ WEB

Detalls del desenvolupament de l'aplicació web.
Front-End amb Flask [6].
Back-End amb FastAPI [5].

VI. 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.
Tot i que oficialment la durada del projecte és de 22 setmanes, per realitzar el projecte han estat 25 setmanes, ja que s'ha començat amb 3 setmanes d'antelació.

VI-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.

VI-A1. Durada de les fases

: Tal i com s'ha esmentat anteriorment, la durada del projecte és de 25 setmanes, sumant les 3 setmanes prèvies de preparació. Per tant la data inicial del projecte ha estat durant la setmana del 29 de gener de 2024.
La fase d'**inici del TFG** correspont a 3 setmanes del total del projecte, que comença el 29 de gener del 2024 i acaba el 18 de febrer de 2024.
La fase de **generació de perfils sintètics** comença el 19 de febrer del 2024 i acaba el 14 d'abril del 2024.

La fase d'**entrenament de la IA** comença el 15 d'abril de 2024 i acaba el 5 de maig del 2024.
La fase de l'**aplicació web** comença el 6 de maig del 2024 i acaba el 26 de maig del 2024.
La fase de la **proposta informe final** comença el 27 de maig del 2024 i acaba el 16 de juny del 2024.
La fase de la **proposta de presentació** comença el 17 de juny del 2024 i acaba el 29 de juny del 2024.
La fase del **lliurament del dossier** comença l'1 de juliol del 2024 i acaba el mateix dia, es a dir, acaba l'1 de juliol del 2024.
La fase del **lliurament del pòster** comença el 2 de juliol del 2024 i acaba el 4 de juliol del 2024.
I l'última fase, **defensa del TFG**, comença el 5 de juliol i acaba el 20 de juliol.

VI-B. Milestones

Durant el transcurs del projecte, s'han determinat les següents *milestones*, per tal de portar un seguiment del projecte i validar que avança de forma satisfactòria. Les *milestones* són:

- **Entrevista inicial:** Primer contacte amb el tutor, per rebre la informació sobre el projecte a realitzar.
- **Informe lliurament inicial:** Realització d'un primer informe, on s'indica els objectius del projecte, la metodologia a utilitzar i la planificació d'aquest projecte.
- **Informe de seguiment I:** Expansió de l'informe inicial amb el treball realitzat fins el moment.
- **Informe de seguiment II:** Redacció de l'informe de desenvolupament del treball ja finalitzat, comprèn tota la informació dels informes anteriors, amb l'afegit del treball realitzat fins a acabr-lo.
- **Proposta informe final:** Entrega de la proposta de l'informe a entregar per a ser evaluat. Ha d'incloure una introducció del prohecte, una explicació dels conceptes que s'utilitzaran durant el desenvolupament d'aquest, el treball realitzat amb els seus resultats i les conclusions finals.
- **Proposta de presentació:** Proposta de la presentació a realitzar durant la defensa del projecte per validar la seva execució.
- **Lliurament del dossier:** Entrega de tota la documentació relacionada amb el projecte.
- **Lliurament del pòster:** Entrega del pòster del projecte, que posteriorment serà posat a cuncurs.
- **Defensa del TFG:** Defensa oral del projecte davant un jurat.

De les *milestones* mencionades, s'han completat amb èxit les tres primeres. Per tant, les *milestones* completades són: **entrevista inicial**, **informe lliurament inicial** e **informe de seguiment I**.

VII. References

Referencias

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

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

[3] 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.

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

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

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

[7] "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/.

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

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

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

[11] "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.

[12] 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.

[13] "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