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. En resum, 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.
- **Paraula 2**: Definició paraula 2.
- **Paraula n**: Definició paraula n.

II. QUE ÉS LA IA?

Definició DE IA.

II-A. Orígens

Orígens de la IA.

II-B. Estat de l'art

Estat actual de la IA.

II-C. Projecció a futur de la IA

Potencial de la tecnologia.

III. GENERACIÓ DE DADES SINTÈTIQUES

Generació de perfils sintètics de persones, basades en paràmetres introduïts en el codi de la generació dels perfils.

III-A. Format d'emmagatzematge de les dades

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: (explicar estructura).
- XML: (explicar estructura).
- JSON (semi estructurada): (explicar estructura).
- Etc.

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: (explicar estructura).
- Imatge: (explicar estructura).
- So: (explicar estructura).
- Vídeo: (explicar estructura).

En el nostre cas les dades generades de forma sintètica, s'emmagatzemaran en el format estructurat **X**.

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 pels diferents algoritmes supervisats que disposa la llibreria Scikit-learn, que són: [4]

- *Linear Models*: (caracterírtiques).
- *Linear and Quadrantic Discriminant Analysis*: (caracterírtiques).
- *Kernel ridge regression*: (caracterírtiques).
- *Support Vector Machines*: (caracterírtiques).
- *Stochastic Gradient Descent*: (caracterírtiques).
- *Nearest Neighbors*: (caracterírtiques).
- *Gaussian Processes*: (caracterírtiques).
- *Cross decomposition*: (caracterírtiques).
- *Naive Bayes*: (caracterírtiques).
- *Decision Trees*: (caracterírtiques).
- *Ensembles: Gradient boosting, random forests, bagging, voting, stacking*: (caracterírtiques).
- *Multiclass and multioutput algorithms*: (caracterírtiques).
- *Feature selection*: (caracterírtiques).
- *Semi-supervised learning*: (caracterírtiques).
- *Isotonic regression*: (caracterírtiques).
- *Probability calibration*: (caracterírtiques).
- *Neural network models (supervised)*: (caracterírtiques).

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

IV-A2. Linear and Quadrantic Discriminant Analysis: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

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

IV-A4. Support Vector Machines: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A5. Stochastic Gradient Descent: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A6. Nearest Neighbors: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A7. Gaussian Processes: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A8. Cross decomposition: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A9. Naive Bayes: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A10. Decision Trees: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A11. 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-A12. Multiclass and multioutput algorithms: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A13. Feature selection: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A14. Semi-supervised learning: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A15. Isotonic regression: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A16. Probability calibration: L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

IV-A17. Neural network models (supervised): L'aplicació de l'algoritme amb el dataset de proves ha tingut un encert del **x %**.

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

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.

- **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.

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.

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/.