Generació de Dades Sintètiques i predicció de dades en entorns reals

Arnau Garriga Riba

Resum — Cada cop s'utilitzen més les dades no reals (dades sintètiques) per a millorar els datasets que s'utilitzen en algorismes de Machine Learning. En aquest treball, s'exploren i s'implementen diferents mètodes o algorismes de dades sintètiques sobre tres datasets basant-se en els seus patrons per a triar, per a cada dataset, el mètode que generi dades sintètiques més semblants a les dades originals. Posteriorement, s'utilitzen algorismes de Machine Learning per a determinar la millora que ha proporcionat la inclusió d'aquestes dades sintètiques en el dataset original. També es realitzen un seguit de visualitzacions i gràfiques amb les quals es pot comparar de manera visual les dades sintètiques generades amb les dades originals.

Paraules clau — Dades sintètiques, Aprenentatge automàtic, Imblearn, Pandas, Sklearn, Visualització de Dades, Python, Xarxes Neuronals, Patrons, Prediccions

Abstract — Non-real data (synthetic data) is increasingly being used to improve the datasets used in Machine Learning algorithms. In this project, different synthetic data methods or algorithms are explored and implemented on three datasets, based on the their patterns to choose, for each dataset, the method that generates synthetic data most similar to the original data. Later, Machine Learning algorithms are used to determine the improvement provided by the inclusion of this synthetic data in the original dataset. A series of visualizations and graphs are also made with which the synthetic data generated can be compared visually with the original data.

Index Terms — Syntehtic Data, Machine Learning, Imblearn, Pandas, Sklearn, Data Visualization, Python, Neural Networks, Patterns, Predictions

1 INTRODUCCIÓ

En enginyeria de dades existeix el concepte de generació de dades sintètiques [*1*], que consisteix en un seguit d'algorismes per incrementar la quantitat de dades, afegint còpies de les dades existents, però amb petites modificacions. Aquestes noves dades sintètiques permeten augmentar la mida i dimensió dels datasets sense perdre l'estructura i el comportament de les originals. Aquest augment de dades millora el rendiment en els models d'aprenentatge automàtics, útils per a molts projectes.

En aquest treball volem aprofundir i investigar l'àmbit de la generació de dades sintètiques.

Les principals aplicacions de la generació de dades sintètiques són les següents:

- **Augmentar les mostres del dataset**: En molts casos, els conjunts de dades disponibles són petits o limitats. Creant dades sintètiques, s'augmenta la quantitat de dades disponibles per obtenir resultats més fiables.
- **Privacitat**: Es generen noves dades per a evitar l'exposició de les dades reals que són sensibles.
- **Reducció de l'overfitting**: Afegint dades sintètiques, es té una millor varietat de dades per entrenar el model i s'aconsegueix que aquest tingui un bon rendiment en les noves dades, és a dir, es redueix l'overfitting.
- **Millora dels resultats en un model de ML**: Al disposar d'una major quantitat de dades, els models de Machine Learning tindran un millor rendiment.
- **Balanceig dels datasets** (que totes les categories tinguin aproximadament la mateixa quantitat de mostres): En molts casos, es necessita tenir un dataset balancejat. Amb les dades sintètiques, es pot augmentar les dades d'una etiqueta (categoria associada a una mostra) amb pocs registres perquè el dataset estigui balancejat.

Afegir dades sintètiques ben generades i rellevants millora significativament la precisió del model, mentre que afegir dades sintètiques inexactes o irrellevants pot tenir efectes negatius. Per això, és important trobar, per a cada dataset, un mètode que li generi correctament dades sintètiques.

Actualment, hi ha diferents algorismes o mètodes de generació de dades sintètiques. Els més importants són:

- Les GANs [GAN] (Xarxes Generatives Antagòniques) generen dades sintètiques mitjançant xarxes neuronals. Es defineixen dues xarxes neuronals: una generadora que crea dades sintètiques i una discriminadora que avalua l'autenticitat d'aquestes dades. Les xarxes s'entrenen juntes en un procés iteratiu de "lluita" entre el generador i el discriminador, fins que el generador pot produir dades sintètiques d'alta qualitat. S'utilitza bastant en els datasets compostos d'imatges.

- SMOTE [SMT] és una tècnica que genera dades extra segons la columna especificada. Fa la representació de cada registre com un punt a l'espai. Després, crea les dades sintètiques, com a punts entremig de les dades ja creades, interpolant característiques entre la mostra seleccionada i els seus veïns com es mostra en la figura 1.

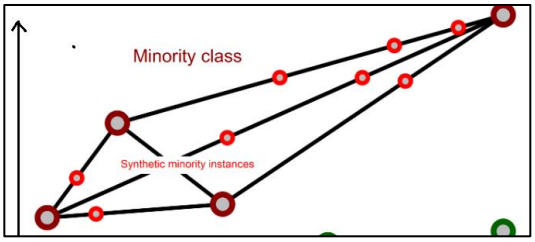

Figura 1. SMOTE. Els 4 punts grans són les dades originals i els 8 punts petits són les dades generades.

SMOTE està implementat en moltes llibreries de ML, com són Imblearn [Imb] (on s'usa força la funció SmoteTomek [ST]), Keras [Ker] o Sklearn [Skl].

- Els VAEs [VAE] són un altre tipus de xarxes neuronals que permeten generar dades sintètiques. Els VAE codifiquen les dades d'entrada en una representació de menor dimensió anomenada espai latent. Utilitzen una capa de codificació per convertir les dades a una distribució a l'espai latent, que sol ser una distribució gaussiana amb mitjana i desviació estàndard. Després es fa la descodificació per generar dades que s'assemblin el més possible a les dades d'entrada originals. Aquest procés es mostra a la figura 2.

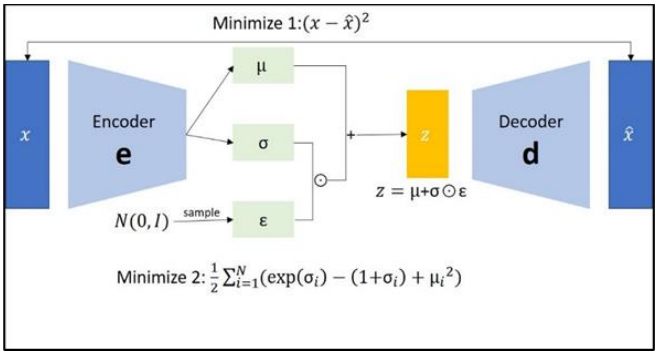

Figura 2. Funcionament d'un Variational Autoencoder.

- L'algorisme de Mescla Gaussiana [GMM] és un model probabilístic que té en compte la distribució de les dades originals per a crear dades artificials.

Veient la taula comparativa, observem que per les 2 xarxes neuronals GAN i VAE, donarien bons resultats datasets grans, amb un alt nombre de registres. Després, GMM aniria bé per datasets amb gran quantitat de variables numèriques. SMOTE és l'algorisme que més s'adaptaria a un dataset comú, estàndard, de característiques i dimensió normals.

Algorisme	Punts forts	Punts febles
GAN / RGA	Útil per datasets tabulars, d'imatges, text, so...	Cal un conjunt d'entrenament gran
SMOTE	Datasets amb les etiquetes desbalancejades	Outliers i soroll
VAE	Dades amb una distribució de probabilitat entre elles	Cal un conjunt d'entrenament gran
GMM	Dades amb una distribució semblant a les dades originals	Dificultat per a relacionar variables categòriques i numèriques

Taula 1. Comparativa dels diferents algorismes.

Un concepte similar al de les dades sintètiques i que, en ocasions, es confon és el de Data Augmentation [2]. És una tècnica que també consisteix a crear dades addicionals sobre un dataset per aconseguir una major diversitat al conjunt d'entrenament. La diferència entre Data Augmentation i la generació de dades sintètiques és que el primer modifica les dades existents per crear noves variacions, com ara rotar, canviar l'escala o aplicar filtres, mentre que la generació de dades sintètiques crea dades completament noves en base als patrons i estructura de les dades originals. Usualment, el terme de Data Augmentation s'acostuma a utilitzar en els datasets d'imatges mentre que el de dades sintètiques s'utilitza en tot tipus de datasets.

Ambdues tècniques es poden utilitzar per millorar el rendiment dels models d'aprenentatge automàtic augmentant la quantitat de dades d'entrenament disponibles.

Hem vist que hi ha diferents algorismes o mètodes per a crear dades sintètiques, i cada mètode genera noves dades de forma diferent, amb propietats diferents. Per tant, donat un dataset, pot ser interessant trobar el millor algorisme segons uns indicadors.

Per tot això, el principal objectiu del projecte és definir criteris per a trobar el millor algorisme per a la creació de dades sintètiques donats datasets de diferents àmbits.

Per a decidir quin és el millor algorisme, s'aplicarà un model de Machine Learning a cada mètode, i s'estudiaran els seus resultats segons un conjunt de mètriques per a veure quin algorisme és millor.

Principals tasques a realitzar:

- Implementar diferents mètodes de generació de dades sintètiques.

- *E-mail de contacte: agarrigariba@gmail.com*
- *Treball tutoritzat per: Enric Martí Gòdia (Departament de Ciències de la Computació)*
- *Curs 2022/23*

- Trobar mètriques adients sobre cada mètode i dataset aplicat, segons la semblança amb les dades originals.
- Experimentar amb diferents mètodes de Machine Learning per fer prediccions, comparant els resultats amb dades sintètiques i sense.
- Implementar visualitzacions que validin els resultats obtinguts.

Per tal de treballar amb les dades sintètiques, s'ha decidit utilitzar diferents eines. L'entorn Pycharm [Pyc], per treballar amb els diferents scripts Python. Google Colab [GoC], per els scripts que inclouen xarxes neuronals. RStudio [RSt], per fer les diferents representacions amb R i finalment ModernCSV [MCSV] per a tractar més fàcilment tots els fitxers csv del treball.

2 DATASETS

Per a provar els algorismes de dades sintètiques hem agafat 3 datasets. En el moment de seleccionar-los, s'ha tingut en compte que siguin de diferents tipologies (esports, medicina, entreteniment i economia) i que tinguin una proporció diferent de variables categòriques i numèriques per a provar les fortaleses o febleses dels algoritmes. Els datasets són:

1. **Malalties cardiovasculars** [KCaDi]: Dataset mèdic que conté informació dels factors de risc per a les malalties cardiovasculars. Com a aspecte positiu, conté tant variables categòriques com numèriques. Les categòriques inclouen pocs valors diferents. Conté 70000 registres amb 14 camps (contant l'índex). Cada registre és una persona adulta.
2. **Netflix** [KNe]: Conté informació de les millors pel·lícules de la plataforma Netflix. Conté una variable d'opinions d'usuaris, que és el que el fa diferent dels altres datasets escollits. Les variables estan bastant correlacionades entre elles. Conté 387 registres amb 8 camps. Cada registre és una pel·lícula.
3. **Futbol** [KFoTr]: Dataset amb informació de tots els fitxatges de futbol fets a tot el món durant l'estiu del 2022. Conté 33625 registres amb 11 camps. Hi ha una àmplia varietat de possibles valors en les seves variables categòriques. Cada registre és un fitxatge.

Prèviament a aplicar els alorismes farem Data Massaging sobre cada dataset. Ho expliquem a continuació.

2.1 Dataset malalties cardiovasculars

Per aquest dataset fem les següents operacions:

- Per no tenir 2 columnes d'identificació del registre, fem que la columna "id" sigui l'índex i esborrem la columna "index".
- Passem la columna "age" de dies a anys perquè quedi més fàcil d'entendre. L'any serà un enter i s'aproximarà sempre cap a baix. Tindrem un rang d'edats d'entre 29 a 64 anys.

Posteriorment, fem una eliminació de registres, que, tenint en compte el que signifiquen, són impossibles. Són dades irreals i, per tant, les esborrarem. Eliminem els casos on:

- El pes (weight) és menor de 40 kilograms.
- L'alçada (height) és menor a 1 metre.
- El seu IMC [3] és inferior a 16 o superior a 90.
- La pressió arterial sistòlica (ap_hi) és inferior a 0 o superior a 220.
- La pressió arterial diastòlica (ap_lo) és inferior a 0 o superior a 140.

Després d'aquests canvis, resta una quantitat de 68829 registres, un 98,33% dels registres inicials i 12 variables.

2.2 Dataset Netflix

Per a aquest dataset creem noves columnes, donat que en té poques i així la creació de dades sintètiques tindrà més joc. Hem decidit crear una columna numèrica i una categòrica, per ampliar columnes dels 2 tipus:

- **Oscars**: Variable numèrica que representa la quantitat de premis Òscar que va rebre la pel·lícula. Per aquesta columna, hem extret la informació de la wikipedia [wiki] utilitzant la llibreria BeautifulSoup [BeS], una llibreria per extreure contingut en format HTML.
- **Original Language**: Variable categòrica que representa la llengua original del film. Per a obtenir aquesta informació hem entrat a l'enllaç de la wikipedia de cada film, extraient l'idioma a la Infobox de la pàgina. Per 121 casos ho hem hagut de fer manualment, doncs la pàgina no tenia Infobox o amb el títol no es podia accedir a l'article de la pel·lícula.

2.3 Dataset futbol

Aquest dataset té un total de 33626 fitxatges. Aquesta quantitat de registres és massa elevada i dificulta el seu tractament. Així doncs, fem una reducció del nombre de registres i ens quedem amb els fitxatges més importants.

- Comencem suprimint els registres de jugadors que s'han retirat o que han estat pujats al primer equip. Aquests registres estaven a partir de l'índex 27645. Després, hem eliminat tots aquells fitxatges a equips que no són de les 10 grans lligues europees (Anglaterra, Alemanya, França, Espanya, Itàlia, Països Baixos, Portugal, Bèlgica, Rússia i Turquia).
- Posteriorment, hem eliminat 4 columnes que ens compliquen la generació de noves dades. Aquestes són els clubs d'origen i destí i les seves lligues. A canvi, hem utilitzat un dataset de FIFA23 [FIFA23] per fer un join i afegir-hi la columna amb la nacionalitat de cada jugador.
- Després, hem convertit la variable "cost" al tipus float, i li hem canviat el nom a "market_value". Posteriorment, ens hem quedat només amb els 1000 fitxatges amb el valor de mercat més alt.
- Finalment, hem canviat el tipus de la variable "date_of_transfer" a tipus data, per poder treballar amb aquests valors com a dates.

Amb aquests canvis, ha canviat la definició del dataset

passant d'un dataset de tots els fitxatges d'estiu a un dataset dels fitxatges més importants de les 10 grans lligues. Però aquest dataset ens interessa més que l'incial, doncs conté dades de jugadors més importants. Treballem amb aquests 1000 registres.

3 DESENVOLUPAMENT

A continuació, explicarem quins són els algorismes que hem implementat per a generar les dades sintètiques. Com els apliquem a cada un dels datasets i quines mètriques utilitzem per a mesurar el rendiment obtingut amb la generació de les dades sintètiques.

3.1 Algorismes a implementar

Hem triat i implementat aquests quatre algorismes o mètodes degut a la seva varietat en el procés de generar les dades. Els aplicarem als tres datasets analizant-ne els resultats:

- Aleatori
- SMOTE
- GMM
- GAN

3.1.1 Aleatori

Algorisme ideat per mi que consisteix en generar noves dades amb una distribució aleatòria de les dades originals, quedant-nos només amb aquells registres que segueixin el comportament de les originals.

Els passos a seguir són els següents:

1. **Iterar columna a columna.** A cada una es selecciona un valor aleatori entre totes les dades originals d'aquella columna. Un cop iterades totes les columnes, es té un nou registre amb valors aleatoris.
2. **Afegir una puntuació.** Com més patrons o comportaments segueixi el nou registre respecte a les dades originals, millor puntuació obtindrà. Com més punts tingui el registre, més semblança tindrà amb les dades originals. Els punts i patrons variaran segons el dataset.
3. **Definir un llindar de punts**. Si un registre el supera, vol dir que és acceptable i, per tant, l'afegirem. En cas contrari, descartarem aquest registre. Aquest llindar estarà definit perquè accepti aproximadament un 10% dels registres aleatoris generats.

Un avantatge d'aquest algorisme és que és bàsic i fàcil d'entendre. Com a inconvenient tenim que s'ha d'adaptar a cada dataset, doncs s'ha de tenir un previ coneixement del dataset per detectar els patrons i definir els punts per a cada un d'ells.

3.1.2 SMOTE

El segon algorisme està basat en la tècnica de SMOTE, que s'utilitza per abordar el problema de desequilibri de classe, és a dir, les dades estan desproporcionadament distribuïdes entre diferents classes o categories. Aquest mètode genera dades sintètiques de la classe minoritària de la variable a predir. Tot i això, en alguns conjunts de dades no es disposa d'una columna objectiu o columna a predir i, per tant, no hi ha classes minoritàries o majoritàries definides. Per adaptar el mètode a aquests casos, hem realitzat una variació de l'algorisme amb els següents passos:

1. Iterar columna a columna.
2. A cada iteració s'especifica una de les columnes com a objectiu.
3. Es creen les dades sintètiques sobre la columna especificada.

Com més columnes tingui el dataset, menys variarà la freqüència dels diferents valors de les columnes respecte a la de les dades originals. Si es veu que aquesta freqüència varia molt, s'aplica una reducció de les dades creades, per a igualar les freqüències.

El bucle només s'iterarà per columnes categòriques. És a dir, si una columna és numèrica s'ignorarà.

Un avantatge d'aquest algorisme és que les dades són fàcils de crear, doncs hi ha moltes funcions que simplement cridant-les ja genera les noves dades. Com a inconvenient, tenim que poden generar-se dades sintètiques massa semblants a les originals.

3.1.3 GMM

Per a la tercera tècnica, utilitzarem una variació del GMM (Gaussian Mixture Model), un model probabilístic que representa una distribució de probabilitat com una suma ponderada de distribucions gaussianes. S'entrenaran les dades originals per després obtenir unes dades sintètiques que segueixin el mateix patró.

A l'hora de fer la primera implementació hem observat que donava resultats bastant aleatoris i, per tant, dolents. Després d'analitzar-ho, hem detectat que es deu al fet que tracta alhora dades categòriques i dades numèriques. Per a millorar-ho, hem adaptat el mètode de la següent manera:

1. Fem ús de la classe LabelEncoder de Sklearn per a codificar les variables categòriques.
2. Fem un escalat de les dades en valors que es mouen entre -10 i 10 perquè totes les característiques tinguin una magnitud comparable.
3. Creem un model GMM per a les dades categòriques i un per a les dades numèriques.
4. Amb les dades sintètiques generades a cada model, creem totes les possibles combinacions.
5. Per a cada possible combinació calcularem la "mean" que serà la distància respecte a les dades actuals.
6. Aquelles combinacions amb la "mean" més baixa seran afegides com a dades sintètiques.

3.1.4 GAN

Com a últim algorisme, utilitzarem una xarxa neuronal GAN (Generative Adversarial Networks), utilitzant la llibreria TensorFlow [TeF] i Keras. Les GANs estan com-

postes de 2 xarxes antagòniques, la generadora i la discriminadora.

Les GANs són molt usades per a la generació de dades sintètiques en datasets d'imatges, però també s'usen en alguns casos en datasets tabulars (dataset estructurat en forma de taula). Per aquests casos, es fa servir la llibreria TabGan [TaG], que està especialitzada en generar dades sintètiques sobre datasets tabulars. L'algorisme segueix els següents passos:

1. Construir una xarxa neuronal amb capes denses que utilitzen la funció d'activació RELU [4].
2. Entrenar el model utilitzant les dades originals.
3. Utilitzar una funció de Tabgan per obtenir les dades sintètiques.

Com a avantatge, les GAN poden produir dades sintètiques d'alta qualitat i varietat. Com a inconvenient, aquestes xarxes són costoses d'entrenar i ajustar.

3.2 Aplicació dels algorismes

Com a resultat d'aplicar els 4 algorismes per a cada dataset hem obtingut 12 grups de dades sintètiques diferents.

S'han generat un 75% de dades sintètiques sobre el total per a cada un dels datasets. Tot i això, en el dataset més gran, el mèdic, pels algorismes de Random Sampling, GMM i GAN, hem hagut d'aplicar una reducció diferent al dataset original doncs donava problemes en espai i temps, que ho feien inviable. Això es pot observar a la taula 2.

	Reducció	*Nous registres*
Aleatori	*20%*	*10325*
SMOTE	*75%*	*51621*
GMM ***GAN***	*2%*	*1032*

Taula 2. Generació de dades sobre el dataset mèdic

Un cop generats els grups de dades sintètiques, s'avaluen amb models de Sklearn diferents.

Per a utilitzar els algorismes, necessitem definir una variable objectiu o target, que és la més representativa d'un dataset i que serà important per poder avaluar les dades sintètiques generades.

Perquè els resultats tinguin més robustesa, hem decidit escollir 2 variables target per a cada dataset. La primera és la que hem considerat com a més representativa. La segona l'hem seleccionada utilitzant una PCA [5] per veure quina és la variable amb el coeficient de correlació més alt.

El que faran els models serà predir aquestes variables per a les dades sintètiques. Com millor es facin les prediccions, millor seran les dades generades. En el nostre cas, escollirem les variables *targets* següents:

- *Cardio* i *Weight* per al dataset de malalties cardiovasculars.
- *Main_production* i *Number_of_votes* per al dataset de Netflix.
- *Market_value* i *Date_of_transfer* per al dataset de futbol.

Podem observar que per a cada dataset, una de les variables triada és categòrica i l'altra numèrica.

Per a predir aquestes variables, emprarem els següents models:

- **Decision Tree** [DT]: Utilitza una estructura d'arbre. Cada node intern de l'arbre representa una decisió basada en un valor d'una variable, mentre que cada node fulla representa una etiqueta o valor numèric.
- **Random Forest** [RF]: Crea múltiples Decision Trees, cada un amb diferents grups de dades. L'output serà la moda (en cas de classificació) o la mitjana (en cas de regressió) de les sortides de cada arbre individual.

Hem triat aquests models perquè ambdós poden gestionar tant variables categòriques com numèriques.

Per a cada dataset, agafarem les dades originals com a dades d'entrenament i els 4 grups de dades sintètiques generades per cada algorisme com a dades de validació. Utilitzarem els 2 models per predir les "variables objectiu" de cada un dels 4 conjunts de dades sintètiques. Aleshores, les dades que han estat més ben predites són les que obtindran millors mètriques. Utilitzarem 4 mètriques diferents per veure el rendiment de cada grup de dades sintètiques.

Si la variable target és categòrica, utilitzarem 2 mètriques:

- **Accuracy**: Prediccions correctes respecte del total. És la mètrica més utilitzada. Millor rendiment com més alta és l'accuracy.
- **F1-Score**: Combina la precisió (proporció de positius identificats que sí són positius) i el recall (proporció de positius respecte a tots els positius que existeixen). Millor rendiment com més alt és la mètrica.

Si la variable target és numèrica, utilitzarem 2 mètriques:

- **MAE** (Mean Aboslute Error): Mesura d'avaluació del rendiment que calcula la mitjana de les diferències absolutes entre els valors predits i els valors reals. Millor rendiment com més baix és el MAE.
- **R squared** (Coeficient de determinació): Valor entre 0 i 1, on 1 indica que el model s'ajusta perfectament a les dades i 0 indica que el model no explica res de la variabilitat de la variable objectiu. Millor rendiment com més proper al 1 és la mètrica.

4 RESULTATS

En aquesta secció mostrem els resultats obtinguts:

Per a cada dataset, mostrem una figura o hi ha els resultats de les 2 variables objectiu. Cada variable té 4 proves o resultats. Subratllem en verd el millor resultat, en vermell el pitjor, en groc si s'acosta molt al millor i en taronja si s'acosta molt al pitjor.

4.1 Resultats dataset malalties cardiovasculars

Amb la predicció de les etiquetes de les variables "Cardio" i "Weight", hem obtingut els resultats mostrats en la figura 3.

Target: cardio

	Algoritme	Accuracy DT Classifier	F1 DT Classifier	Accuracy RF Classifier	F1 RF Classifier
0	Aleatori	0.501938	0.460084	0.501938	0.400932
1	SMOTE	0.858527	0.856016	0.889535	0.888889
2	GMM	0.540698	0.496815	0.594961	0.576923
3	GAN	0.845930	0.829582	0.843992	0.833161

Target: weight

	Algoritme	MAE DT Regression	R2 DT Regression	MAE RF Regression	R2 RF Regression
0	Aleatori	14.792205	-1.901456	9.652615	-0.257521
1	SMOTE	6.129797	0.368729	6.830295	0.556743
2	GMM	12.396878	-1.022243	9.629759	-0.155910
3	GAN	6.951590	0.120411	7.418733	0.368147

Figura 3. Mètriques obtingudes en el dataset mèdic

Veiem que, l'algorisme de SMOTE és el que produeix els millors valors (subratllats en verd), és a dir, dades sintètiques més semblants a les dades originals. El segon millor és el GAN amb seguit pel GMM i l'aleatori. En aquest cas, SMOTE obté els millors resultats en les 8 proves.

4.2 Resultats dataset netflix

Després d'haver provat els models al dataset de Netflix predient les variables "MAIN_PRODUCTION" i "NUMBER_OF_VOTES" mostrem els resultats en la figura 4.

Target: MAIN_PRODUCTION

	Algoritme	Accuracy DT Classifier	F1 DT Classifier	Accuracy RF Classifier	F1 RF Classifier
0	Aleatori	0.192440	0.221378	0.027491	0.047261
1	SMOTE	0.048276	0.070599	0.003448	0.000088
2	GMM	0.010345	0.010142	0.006897	0.001349
3	GAN	0.051724	0.071715	0.241379	0.236815

Target: NUMBER_OF_VOTES

	Algoritme	MAE DT Regression	R2 DT Regression	MAE RF Regression	R2 RF Regression
0	Aleatori	92453.068729	-0.277475	92462.472612	-0.277532
1	SMOTE	122343.837931	-0.557347	122346.262069	-0.557411
2	GMM	129379.251724	-0.589943	129371.686759	-0.589882
3	GAN	107825.472414	-0.511598	107830.302069	-0.511628

Figura 4. Mètriques obtingudes en el dataset de Netflix

Veiem que el millor algorisme és l'aleatori, seguit pel GAN que obté els millors resultats en els classificadors de Random Forest. Després, amb uns resultats més dolents i similars entre ells tenim el SMOTE i el GMM.

4.3 Resultats dataset futbol

En la figura 5 mostrem els resultats d'aplicar els algorismes al dataset de futbol per predir les variables "market_value " i "date_of_transfer".

Target: market_value

	Algoritme	MAE DT Regression	R2 DT Regression	MAE RF Regression	R2 RF Regression
0	Aleatori	9.362461	-7.328058	6.162308	-0.998587
1	SMOTE	5.390917	-1.543998	4.182808	0.120948
2	GMM	8.356461	-4.646945	6.244108	-0.817251
3	GAN	10.005688	-0.124037	9.651341	0.056126

Target: date_of_transfer

	Algoritme	Accuracy DT Classifier	F1 DT Classifier	Accuracy RF Classifier	F1 RF Classifier
0	Aleatori	0.083223	0.085830	0.243065	0.145100
1	SMOTE	0.289683	0.272846	0.387566	0.362645
2	GMM	0.105727	0.111810	0.268722	0.158056
3	GAN	0.125661	0.134060	0.313492	0.236463

Figura 5. Mètriques obtingudes en el dataset de futbol

Podem observar que l'algorisme SMOTE és el que produeix millors resultats, sobretot a l'hora de predir la variable *date_of_transfer*. El segon millor el GMM, el tercer el GAN i l'últim l'aleatori.

El fet que l'algorisme aleatori sigui el millor per al dataset de Netflix creiem que es deu al fet que conté informació d'opinions subjectives i basades en les preferències o avaluacions dels usuaris. Aquest fet, és el que el diferencia dels altres datasets, amb dades més objectives.

5 ANÀLISIS DELS RESULTATS

S'han fet un seguit de prediccions amb Machine Learning i visualitzacions comparant les dades originals amb les noves per a validar els resultats del millor model a cada dataset.

En els següents subapartats veurem aquestes prediccions i visualitzacions.

5.1 Prediccions

Per a cada dataset agafarem les dades sintètiques que han donat millors resultats i farem 2 iteracions del codi: una utilitzant només les dades originals, i l'altre, utilitzant les dades originals + les dades sintètiques seleccionades. Compararem els resultats de les 2 execucions per veure l'efecte d'aquestes dades sintètiques en el rendiment dels models.

En aquestes execucions s'utilitzaran els mateixos algorismes (Decision Tree i Random Forest) i les mateixes mètriques que en l'apartat anterior.

També, en comptes d'utilitzar un 75% de dades sintètiques respecte a les originals, tornarem a generar dades sintètiques, ara només amb el millor algorisme, creant-ne un 500% més perquè és pugui veure millor la diferència. En la figura 6 es mostra aquest procés.

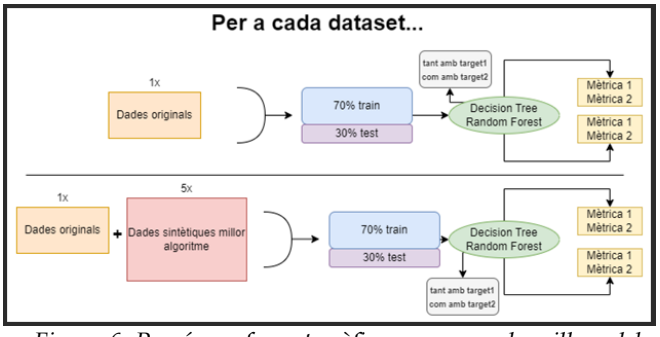

Figura 6. Procés, en format gràfic, per a veure la millora dels resultats

Els resultats obtinguts per a cada dataset després d'aquests processos es mostren a continuació.

Per al dataset mèdic, en el qual hem utilitzat dades generades amb SMOTE, tenim:

- Agafant "cardio" com a target, s'ha obtingut unes millores d'entorn el 140%, passant d'una accuracy o F1 del 60% fins al 90%.
- Agafant "weight" com a variable target, els valors de MAE s'han reduït i els de R2 han augmentat, cosa que indica una millora en la precisió i ajustament del model.

En el cas del dataset de Netflix, en el que hem emprat

dades generades amb l'algorisme aleatori tenim:

- Agafant la variable "MAIN_PRODUCTION" com a target, els resultats han millorat, però aquesta millora no ha estat tan notable, ha estat d'un 112%.
- Agafant la variable "NUMBER_OF_VOTES" com a target, hi ha hagut una lleugera millora dels resultats.

Per al dataset de futbol hem utilitzat les dades generades amb SMOTE. Els resultats obtinguts són els següents:

- Agafant "market_value" com a target, els valors de MAE s'han reduït i els R2 augmentat, cosa que indica una millora en el rendiment del model.
- Agafant "date_of_transfer" com a target, hi ha hagut una millora en el rendiment dels models utilitzant les dades sintètiques en comparació amb només les dades originals. S'ha passat d'una accuracy/recall d'entorn al 20% al 45%.

Seguidament, veurem més clarament aquests resultats amb visualitzacions.

5.2 Visualitzacions

En aquest apartat es fan visualitzacions en les quals es poden comparar les dades originals amb les dades sintètiques generades pel millor algorisme per a cada dataset utilitzat.

5.2.1. Visualitzacions sobre el dataset de malalties cardiovasculars

Per a aquest dataset, hem utilitzat les dades sintètiques generades amb la tècnica SMOTE. Les gràfiques es mostren en les figures 7 i 8.

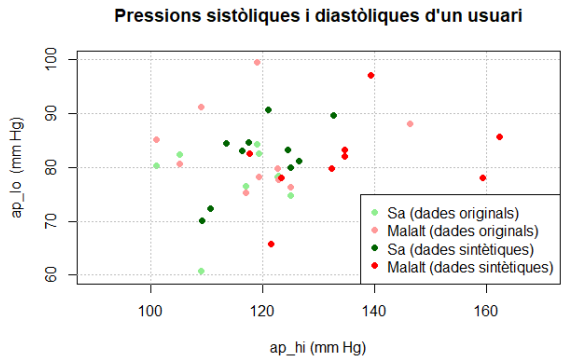

Figura 7. Gràfica de punts generada amb R de les dades de malalties cardiovasculars

En la gràfica de la figura 7 mostrem la relació entre la pressió sistòlica i diastòlica de les dades. Hi ha 4 grups de dades diferents, segons si l'usuari està malalt o sa (variable cardio=1) o si les dades són originals o sintètiques.

Observant la posició i color dels punts, veiem que tant les dades originals com les sintètiques es comporten de la mateixa manera, és a dir, quan l'usuari és sa tendeixen a un ap_hi i ap_lo baix, i a la inversa.

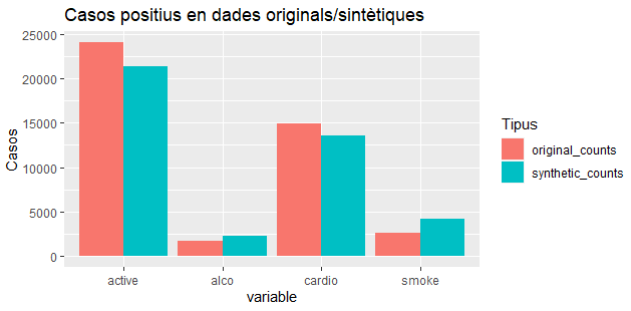

Figura 8. Gràfica de barres generada amb R de les dades de malalties cardiovasculars

A la gràfica de la figura 9 es pot veure la quantitat de dades on aquestes variables són positives. Observem que, tant les dades originals com les sintètiques s'obtenen resultats similars.

5.2.2. Visualitzacions sobre el dataset de Netflix

Hem utilitzat les dades sintètiques generades amb la tècnica aleatòria per al dataset de Netflix. Les gràfiques creades es mostren en les figures 9 i 10.

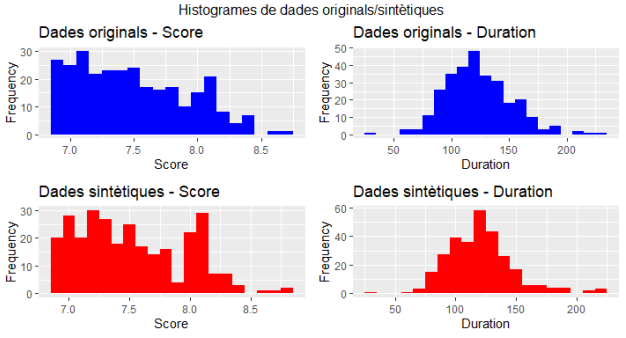

Figura 9. Histogrames generats amb R de les dades de Netflix

En la figura 9 mostrem els histogrames de 2 variables categòriques del dataset de Netflix (Score i Duration). Observem que per les 2 variables, les dades originals (histograma blau) i les sintètiques (histograma vermell) produeixen histogrames molt similars.

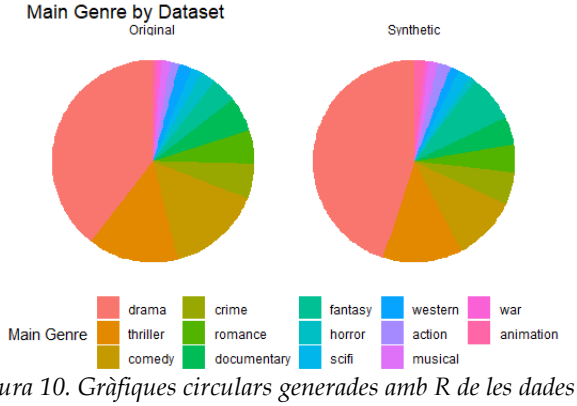

Figura 10. Gràfiques circulars generades amb R de les dades de Netflix

La figura 10 mostra la freqüència dels diferents possibles gèneres. Com podem observar, les dades sintètiques segueixen la freqüència de les dades originals, amb Drama, Thriller i Comèdia com a principals.

5.2.3. Visualitzacions sobre el dataset de futbol

Per al conjunt de dades de futbol, hem utilitzat les dades sintètiques generades mitjançant la tècnica SMOTE. Mostrem les gràfiques en les figures 11 i 12.

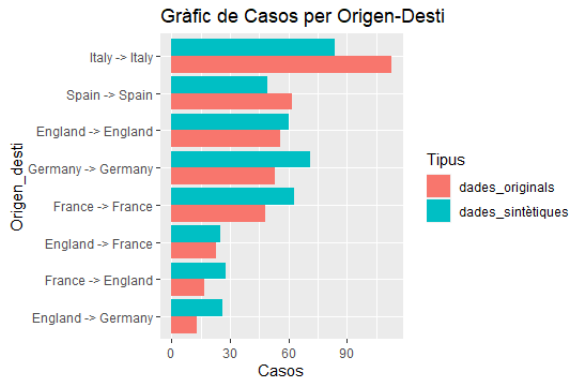

Figura 11. Gràfica de barres generada amb R de les dades de futbol

La figura 11 mostra les combinacions d'origen i destí, més freqüents en els 2 datasets de dades originals i sintètiques. Podem observar que no hi ha grans canvis entre el dataset amb dades originals (vermell) i el dataset amb dades sintètiques (blau).

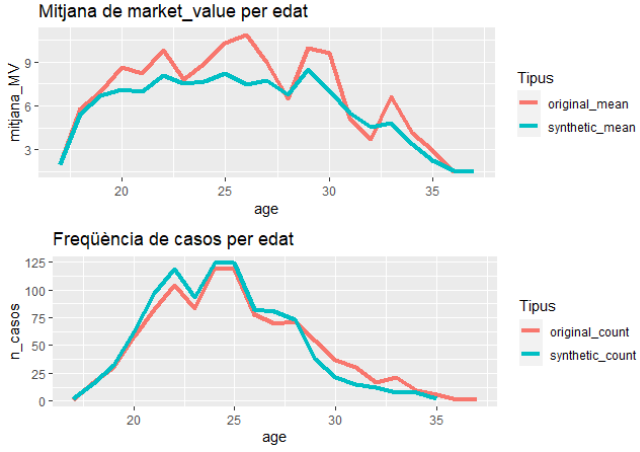

Figura 12. Gràfiques lineals generades amb R de les dades de futbol

A les gràfiques de la figura 12 es mostren les freqüències següents:

- Valor de mercat segons l'edat
- Nombre de casos segons l'edat.

Observem que tant les dades originals com les sintètiques tenen una forma de línia o comportament, semblant.

6 CONCLUSIONS I MILLORES

- S'han implementat 4 algorismes de generació de dades sintètiques amb algunes variacions per millorar el rendiment i la qualitat.
- S'ha aconseguit triar un mètode òptim per a cada dataset basat en certes mètriques.
- Aplicant les dades sintètiques generades amb aquests algorismes, s'ha produït una millora notable en els resultats del models de Machine Learning implementats.
- S'han aplicat els algorismes sobre 3 datasets i comparat els resultats segons unes mètriques per a trobar el millor algorisme per a cada dataset.
- S'ha comprovat la millora obtinguda a cada dataset utilitzant prediccions.
- S'ha validat la generació correcta de les dades sintètiques gràcies a un conjunt de visualitzacions.

Pel que fa a les millores:

- Provar algun algorisme de generació de dades sintètiques més per a tenir més opcions.
- Afegir un quart dataset d'una altra tipologia per a proporcionar una visió més completa i precisa del rendiment de l'algorisme o mètode utilitzat.
- Utilitzar un mètode addicional per calcular com de bé es generen dades sintètiques. Amb això, tenim una mesura addicional de la qualitat de les dades sintètiques generades.
- Utilitzar una eina de visualització més avançada com Tableau, Power BI o Qlik, que ens permet fer gràfiques més elaborades.

AGRAÏMENTS

Vull expressar el meu sincer agraïment al meu tutor, el Sr. Enric Martí, per tot el seu suport i orientació durant el treball. Agraeixo la seva paciència, les seves explicacions detallades i la seva disponibilitat per respondre totes les meves preguntes i dubtes.

BIBLIOGRAFIA

[BeS] https://beautiful-soup-4.readthedocs.io/en/latest/, documentació de la llibreria BeautifulSoup (Data últim accés: març del 2023)

[DT] https://scikit-learn.org/stable/modules/tree.html, explicació del classificador i regressor de Decision Tree (Data últim accés: maig del 2023)

[FIFA23] https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset, enllaç al dataset de FIFA23 (Data últim accés: abril del 2023)

[GAN] https://arxiv.org/abs/1406.2661, informe presentat per Ian Goodfellow i el seu equip l'any 2014 sobre les GANs (Data últim accés: abril del 2023)

[GMM] https://scikit-learn.org/stable/modules/mixture.html, Gaussian Mixture Models amb Sklearn (Data últim accés: abril del 2023)

[GoC] https://colab.research.google.com/, plataforma online de Google on es pot executar codi amb CTU o GPU. (Data últim accés: abril del 2023)

[Imb] https://imbalanced-learn.org/stable/, pàgina amb la documentació de la llibreria imblearn (Data últim accés: març del 2023)

[KCaDi] https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas, lloc web on es troba el dataset mèdic (Data últim accés: febrer del 2023)

[Ker] https://keras.io/, pàgina amb la documentació de la biblioteca de xarxes neuronals Keras (Data últim accés: març del 2023)

[KFoTr] https://www.kaggle.com/datasets/mohamedelhefenawy/summer-football-transfers, enllaç del Kaggle amb el datasets de fitxatges de futbol (Data últim accés: febrer del 2023)

[KNe] https://www.kaggle.com/datasets/thedevastator/netflix-top-rated-movies-and-tv-shows-2020-2022?select=Best+Movies+Netflix.csv, apartat de Kaggle amb el dataset de Netflix (Data últim accés: febrer del 2023)

[MCSV] https://www.moderncsv.com/documentation/, documentació de l'eina ModernCSV (Data últim accés: maig del 2023)

[Pyc] https://www.jetbrains.com/pycharm/, pàgina principal de l'entorn de desenvolupament, Pycharm (Data últim accés: maig del 2023)

[RF] https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, explicació del classificador Random Forest (Data últim accés: abril del 2023)

[RSt] https://posit.co/download/rstudio-desktop/, documentació de l'entorn de programació en R, RStudio

[Skl] https://scikit-learn.org/stable/, documentació de l'extensió Scikit-Learn, també conegut com sklearn (Data últim accés: febrer del 2023)

[SMT] https://ieeexplore.ieee.org/abstract/document/1026736, article inicial de SMOTE publicat al 2002 (Data últim accés: abril del 2023)

[ST] https://imbalancedlearn.org/stable/references/generated/imblearn.combine.SMOTETomek.html, documentació del mètode SMOTETomek (Data últim accés: febrer del 2023)

[TaG] https://pypi.org/project/tabgan/, guia de la llibreria Tabgan (Data últim accés: abril del 2023)

[TeF] https://www.tensorflow.org/, pàgina oficial de la biblioteca de Tensorflow (Data últim accés: abril del 2023)

[VAE] https://arxiv.org/abs/1312.6114, article inical de Diederik Kingma i Max Welling amb la explicació dels VAE (Data últim accés: abril del 2023)

[wiki] https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films, taula amb totes les pel·lícules premiades als Òscar. (Data últim accés: març del 2023)

APÈNDIX

A1. GLOSSARI

1. Dades sintètiques: Les dades sintètiques consisteixen en la generació artificial de dades que imiten les característiques i patrons de les dades reals. Aquestes dades són creades utilitzant algoritmes i tècniques específiques amb el propòsit de simular el comportament i les propietats de les dades existents. Les dades sintètiques són utilitzades en diversos camps, com la investigació científica, l'anàlisi de dades, l'aprenentatge automàtic i la privadesa de dades. La generació de dades sintètiques permet treballar amb conjunts de dades més àmplies i diverses, facilitant l'exploració, el desenvolupament i l'avaluació de models i algorismes sense comprometre la privadesa o la disponibilitat de les dades reals.
2. Data Augmentation: Tècnica utilitzada en el processament de dades a l'àmbit del Machine Learning. Consisteix a generar noves instàncies de dades a partir de les existents mitjançant l'aplicació de transformacions com ara rotacions, translacions, canvis d'escala o modificacions de color. L'objectiu principal és augmentar la quantitat i la diversitat de les dades d'entrenament, cosa que pot millorar el rendiment dels models.
3. IMC (Index de Massa Corporal (IMC): Mesura per evaluar la relació entre el pes i l'altura d'una persona. Es calcula dividint la massa entre l'altura al quadrat. Indica si una persona té baix pes, pes normal o sobrepès.
4. RELU (Rectified Linear Unit): Funció d'activació àmpliament emprada a les xarxes neuronals artificials. Consisteix a retornar el valor d'entrada si és més gran que zero, i zero en cas contrari. La simplicitat i eficiència de la funció RELU la converteixen en una opció popular a moltes aplicacions de Deep Learning.
5. PCA (Principal Component Analysis): Tècnica de reducció de dimensionalitat utilitzada en anàlisi de dades i aprenentatge automàtic. Permet transformar un conjunt de variables correlacionades en un nou conjunt de variables no correlacionades anomenades components principals.