UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Kumar Mahi, Namanmahi. *Generació i avaluació de dades sintètiques tabulars amb SDV*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337355

under the terms of the  license.

Generació i avaluació de dades sintètiques tabulars amb SDV

Namanmahi Kumar Mahi

Tutor: Guillermo Navarro Arribas

14 de juny de 2026 · Curs 2025/26

Resum– Aquest treball compara tres sintetitzadors de dades tabulars de la biblioteca SDV —GaussianCopula, CTGAN i TVAE— sobre el conjunt UCI Bank Marketing en dotze combinacions (dos volums $\times$ dos escenaris $\times$ tres sintetitzadors) i tres eixos: utilitat (*Train-on-Synthetic, Test-on-Real*), privacitat (*Distance to Closest Record* i *Membership Inference Attack*) i fidelitat estadística (SDMetrics). Els resultats confirmen que cap sintetitzador domina simultàniament els tres eixos: GaussianCopula maximitza la privacitat però requereix calibració post-hoc per recuperar la utilitat; TVAE ofereix la millor utilitat però presenta el menor marge de privacitat i un biaix de prevalença en absència de variables molt predictives; i CTGAN amb mostratge condicional és l'únic que controla el *class balance* mantenint un marge de privacitat acceptable. La tria del sintetitzador és, per tant, una decisió sobre el punt desitjat del compromís utilitat–privacitat, no sobre la qualitat absoluta.

Paraules clau– dades sintètiques, dades tabulars, privacitat, GaussianCopula, CTGAN, TVAE, TSTR, DCR, Membership Inference Attack, SDMetrics, calibració de probabilitats.

Abstract– This work compares three synthetic tabular data synthesizers from the SDV library —GaussianCopula, CTGAN and TVAE— on the UCI Bank Marketing dataset across twelve configurations (two volumes $\times$ two scenarios $\times$ three synthesizers), evaluating three axes: downstream utility (Train-on-Synthetic, Test-on-Real with Logistic Regression and Random Forest), privacy (Distance to Closest Record and Membership Inference Attack), and statistical fidelity (SDMetrics). The results confirm that no synthesizer simultaneously dominates all three axes: GaussianCopula maximizes privacy but requires post-hoc calibration to recover utility; TVAE offers the best utility but has the smallest privacy margin and a prevalence bias when highly predictive variables are absent; and CTGAN with conditional sampling is the only architecture that controls class balance while maintaining an acceptable privacy margin. Synthesizer selection is a decision about the desired utility–privacy trade-off point, not about absolute quality.

Keywords– synthetic data, tabular data, privacy, GaussianCopula, CTGAN, TVAE, TSTR, DCR, Membership Inference Attack, SDMetrics, probability calibration.

Agraïments– Vull dedicar unes paraules a les persones que han fet possible aquest camí. Als meus pares, Rakesh Kumar i Rajni Devi, per acompanyar-me de manera incondicional al llarg d'aquests anys de carrera i per ser-hi sempre, tant en els moments bons com en els més difícils. Al meu germà, Prateek Kumar, el meu petit gran pilar: tot i la seva edat, no ha deixat mai d'ajudar-me en la mesura de les seves possibilitats. Als meus amics de la universitat, que han fet aquest recorregut més fàcil, més lleuger i molt més amè. I, finalment, al meu tutor de Treball de Fi de Grau, Guillermo Navarro Arribas, per la guia i les orientacions al llarg de tot el projecte.

1 INTRODUCCIÓ — CONTEXT DEL TREBALL

La compartició de dades és un motor central de la ciència de dades, però sovint topa amb barreres de privacitat: molts conjunts contenen informació personal o comercialment sensible que no es pot difondre lliurement. Les *dades sintètiques* —registres artificials generats per un model que ha après l'estructura estadística de les dades reals— ofereixen una via per superar aquesta barrera: si el conjunt sintètic preserva la utilitat de l'original per a les tasques descendents (*downstream*) i, alhora, no revela informació sobre cap individu real, pot substituir les dades originals en escenaris on aquestes no es poden compartir.

Aquesta promesa, però, amaga una tensió: com més fidelment un sintetitzador reprodueix les dades reals, més útils són les dades sintètiques, però també més gran és el risc que filtri informació sobre els registres d'entrenament. El treball estudia empíricament aquest compromís sobre un cas realista i desbalancejat —el conjunt UCI Bank Marketing [1]— comparant tres arquitectures de síntesi de famílies diferents (una estadística clàssica i dues basades en xarxes neuronals) segons tres eixos: utilitat, privacitat i fidelitat estadística. La hipòtesi de partida és que *cap sintetitzador domina simultàniament en els tres eixos*, i que la tria és, per tant, una decisió contextual sobre el punt desitjat del compromís utilitat–privacitat.

La rellevància pràctica d'aquest problema ha crescut considerablement en els últims anys. Normatives com el Reglament General de Protecció de Dades (GDPR) a Europa o la *Health Insurance Portability and Accountability Act* (HIPAA) als Estats Units imposen restriccions estrictes sobre la compartició de dades personals, elevant el cost de la col·laboració científica i de la innovació basada en dades. Les dades sintètiques han emergit com una resposta tècnica a aquestes restriccions en sectors com la salut —on compartir historials de pacients sense anonimitzar comportaria sancions legals greus—, la banca —on les dades de comportament financer dels clients estan subjectes a regulació sectorial— o l'administració pública. La condició mínima per a l'adopció pràctica d'aquesta alternativa és precisament la que s'estudia en aquest treball: que les dades sintètiques siguin alhora *útils* per a les tasques descendents i *segures* quant a la privacitat dels individus que les originen, i que existeixin mètriques objectives per quantificar ambdues propietats de manera conjunta.

1.1 Objectius

La taula 1 recull els objectius fixats a l'Informe Inicial i desenvolupats al llarg dels seguiments, amb el seu estat final. Els objectius s'articulen en seqüència lògica: O1 i O3 estableixen la infraestructura reproductible i el punt de referència real–real que fan comparables tots els resultats posteriors; O2 implementa el sintetitzador de referència; i O4 i O5 aporten la contribució principal de comparativa i discussió del *trade-off*. Les aportacions A1–A3 van emergir de l'observació empírica i no d'una planificació *a priori*, cosa que garanteix que responen a necessitats reals del treball. Respecte del Seguiment 2, l'abast ha crescut de manera substancial: d'un únic sintetitzador de referència (GaussianCopula) sobre quatre configuracions i una prova pilot de CTGAN, s'ha passat a una comparativa completa de **tres sintetitzadors sobre dotze combinacions**, amb dues mètriques de privacitat i diverses contribucions metodològiques no previstes inicialment (calibració post-hoc, *Membership Inference Attack* i mostratge condicional).

TAULA 1: OBJECTIUS DEL TREBALL I ESTAT FINAL.

Objectiu	**Estat**
O1. Pipeline reproductible PREPARE→TRAIN→EVAL	Completat
O2. Síntesi i avaluació amb GaussianCopula (4 configs)	Completat
O3. Baseline real→real validat (tolerància 0,005 AUC)	Completat
O4. Comparativa amb CTGAN i TVAE (12 combinacions)	Completat
O5. Discussió del *trade-off* utilitat–privacitat–qualitat	Completat
Contribucions addicionals	
A1. Calibració post-hoc del Random Forest (Platt/isotònica)	Completat
A2. Segona mètrica de privacitat (*Membership Inf. Attack*)	Completat
A3. Mostratge condicional de CTGAN per al *class balance*	Completat

1.2 Organització del document

Les seccions 2–6 cobreixen, respectivament, l'estat de l'art, la metodologia, els resultats, la discussió i les conclusions. El diagrama de Gantt i les taules de suport es recullen a l'Apèndix A.6.

2 ESTAT DE L'ART

2.1 Síntesi de dades tabulars

La generació de dades sintètiques tabulars ha evolucionat a partir de dues grans famílies metodològiques, cadascuna amb un enfocament diferent per aprendre com es distribueixen les dades.

La biblioteca SDV. La **Synthetic Data Vault** (SDV) [2, 3] és la biblioteca de codi obert emprada en aquest treball. SDV proporciona una interfície unificada per entrenar, mostrar i avaluar sintetitzadors tabulars, gestionant automàticament els metadades (tipus de variables, distribucions) i les dependències. Integra tres sintetitzadors principals —GaussianCopulaSynthesizer, CTGANSynthesizer i TVAESynthesizer— que corresponen a les dues famílies descrites a continuació.

La primera família, **estadística clàssica**, modela la distribució conjunta descomponent-la en marginals i una estructura de dependència explícita. El fonament teòric és el **teorema de Sklar** [5]: qualsevol distribució conjunta $F(x_1, \ldots, x_d)$ es pot escriure com $F(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$, on C és una *còpula* que captura únicament l'estructura de dependència entre les variables, desacoblada de les seves distribucions marginals F_i. La `GaussianCopulaSynthesizer` de SDV assumeix que C és gaussiana, cosa que la fa lleugera, interpretable i

executable en CPU, però amb capacitat limitada per capturar relacions no lineals o multivariades complexes.

La segona família, **generativa profunda**, aprèn la distribució amb xarxes neuronals de manera implícita.

GAN (Generative Adversarial Network). Les GAN [6] es fonamenten en un joc entre dos components: el **generador** G, que pren soroll aleatori i intenta produir mostres indistingibles de les dades reals; i el **discriminador** D, que intenta distingir les mostres reals de les generades. L'entrenament alterna: D s'entrena a classificar reals vs. sintètics; G s'entrena a enganyar D. L'equilibri de Nash d'aquest joc de suma zero produeix, quan s'assoleix, un generador capaç de mostrejar de la distribució real.

CTGAN (*Conditional Tabular GAN*) [7] adapta el paradigma GAN a dades tabulars. El **generador** pren soroll aleatori i un vector de condició (que codifica el valor d'una categoria seleccionada) i produeix un registre tabular sintètic. El **discriminador** rep un registre (real o sintètic) i ha de determinar si és autèntic. Per tractar variables contínues multimodals —freqüents en dades tabulars— CTGAN aplica una normalització mode-específica basada en una barreja de gaussianes. A més, empra un *mostratge condicional* durant l'entrenament per garantir que el generador aprengui a reproduir la distribució de cada categoria de manera equitativa.

TVAE (*Tabular Variational Autoencoder*) [7, 8] adapta el paradigma VAE a dades tabulars. Un **codificador** comprimeix cada registre real en un vector latent de distribució gaussiana; un **decodificador** reconstrueix registres sintètics mostrejant d'aquest espai latent. L'entrenament minimitza una cota inferior de la log-versemblança (ELBO), que combina la fidelitat de la reconstrucció amb una regularització que manté l'espai latent ordenat. En inferència, se mostreja del prior gaussià i es decodifica per obtenir nous registres. A diferència de CTGAN, no hi ha component adversari: el procés és més estable però pot capturar menys variabilitat.

El camp ha avançat ràpidament en els últims anys. **CTAB-GAN** [19] (Zhao et al., 2021) estén CTGAN amb un codificador condicional especialitzat per a variables d'alta cardinalitat. Els models de **difusió tabulars** [18] (TabDDPM, Kotelnikov et al., 2023) han superat GAN i VAE en múltiples *benchmarks*, si bé a cost computacional notablement superior. Enfocaments basats en *transformers* tabulars [21] amplien la frontera. Totes tres línies queden fora de l'abast d'aquest treball per raons pràctiques: CTAB-GAN no forma part de la biblioteca SDV emprada; TabDDPM i els *transformers* requereixen un cost d'ajust substancialment superior al disponible en l'entorn de Colab gratuït; i la comparació equitativa requeriria un marc d'avaluació diferent. Constitueixen, per tant, el referent natural per a treball futur.

2.2 Marc d'avaluació

No hi ha una mètrica única de qualitat; la pràctica habitual és avaluar les dades sintètiques en tres eixos complementaris. L'eix d'**utilitat** s'avalua amb el protocol *Train-on-Synthetic, Test-on-Real* (TSTR) [9]: s'entrena un model descendent sobre dades sintètiques i es verifica sobre dades reals; com més s'acosta al rendiment del model homòleg entrenat sobre dades reals (TRTR), més útils són les dades sintètiques. En aquest treball s'usen Logistic Regression i Random Forest [10, 11] com a models descendants. L'eix de **privacitat** s'avalua amb dues mètriques complementàries: la *Distance to Closest Record* (DCR), que mesura la proximitat de cada registre sintètic al conjunt d'entrenament i detecta si hi ha còpies pràcticament exactes, i el *Membership Inference Attack* (MIA) [15], que quantifica si un atacant pot determinar si un registre concret ha format part del conjunt d'entrenament. Totes dues són mesures empíriques, a diferència de les garanties formals de la privacitat diferencial [16], que proporciona cotes matemàticament verificables però exigeix modificar el procés d'entrenament i comporta un cost mesurable en utilitat. En la pràctica, un *ratio* DCR superior a la unitat s'empra com a llindar operatiu de seguretat [20]: indica que les dades sintètiques no s'acosten al train més del que ho fan les dades de test, és a dir, no aporten informació addicional sobre els individus més enllà de la que ja conté un *hold-out* independent. Finalment, l'eix de **fidelitat estadística** s'avalua amb SDMetrics [4], que quantifica la semblança via *Column Shapes* (distribucions marginals) i *Column Pair Trends* (dependències bivariades entre parells de variables).

Un repte transversal és el **desbalanceig de classes**. Tècniques clàssiques com SMOTE [17] l'aborden per sobremostratge; en la síntesi, el mostratge condicional de CTGAN [7] permet forçar la proporció de classes desitjada directament en la generació.

2.3 Posicionament del treball

La literatura existent presenta limitacions concretes que aquest treball aborda. La comparativa original de Xu et al. [7] avalua CTGAN i TVAE principalment en utilitat TSTR sobre conjunts equilibrats, sense cap mètrica de privacitat. CTAB-GAN [19] amplia la cobertura de distribucions però tampoc mesura el risc de pertinença. Les revisions generals del camp [20] constaten que molt poques avaluacions combinen els tres eixos simultàniament ni diagnostiquen artefactes específics com la descalibració o el biaix de prevalença en conjunts desequilibrats. Aquest treball aporta una **comparativa controlada i simultània** de tres arquitectures de famílies diferents sobre un cas real i fortament desbalancejat, avaluades en els tres eixos amb **dues** mètriques de privacitat complementàries, amb diagnòstic explícit dels artefactes d'avaluació que distorsionen la lectura de la utilitat quan no es detecten i es gestionen. En concret, la comparativa de tres sintetitzadors sobre dotze combinacions (O4) i la discussió del *trade-off* resultant (O5) cobren els buits identificats en Xu et al. [7] i en [20]; i el diagnòstic del Cas A de descalibració (A1) i el mostratge condicional (A3) són contribucions pràctiques no reportades en comparatives prèvies similars.

3 METODOLOGIA

3.1 Conjunts de dades i escenaris

S'empra el conjunt UCI Bank Marketing [1] (campanyes telefòniques d'un banc portuguès) en dos volums: `bank-additional-full` (41.188 registres, amb deu indicadors socioeconòmics addicionals) i `bank-full` (45.211 registres, sense indicadors macroeconòmics). Ambdós permeten comprovar si les conclusions es mante-

nen independentment de la riquesa informativa. La variable objectiu `y` (11,7 % de positius) fa el conjunt fortament desbalancejat. La taula 2 detalla les principals variables.

TAULA 2: PRINCIPALS VARIABLES DEL CONJUNT UCI BANK MARKETING. LES VARIABLES MARCADES AMB (†) NOMÉS ESTAN PRESENTS EN `BANK-ADDITIONAL-FULL`.

Variable	Tipus	Descripció
Dades del client		
`age`	numèric	Edat del client
`job`	categòric	Tipus de feina (12 cat.)
`marital`	categòric	Estat civil
`education`	categòric	Nivell d'estudis
`default`	categòric	Crèdit en default?
`housing`	categòric	Té hipoteca?
`loan`	categòric	Té préstec personal?
Campanya de contacte		
`contact`	categòric	Canal de contacte
`month`	categòric	Mes de l'últim contacte
`duration`	numèric	Durada (s) de l'última trucada
`campaign`	numèric	Contactes en la campanya actual
`pdays`	numèric	Dies des de l'últim contacte
`poutcome`	categòric	Resultat campanya anterior
Indicadors macroeconòmics (†)		
`emp.var.rate`	numèric	Taxa variació de l'ocupació
`euribor3m`	numèric	Tipus euribor 3 mesos
`nr.employed`	numèric	Nombre de persones ocupades
`cons.price.idx`	numèric	Índex de preus al consum
`cons.conf.idx`	numèric	Índex de confiança del consumidor
Variable objectiu		
`y`	binari	Ha contractat el dipòsit?

Per a cadascun dels dos volums es defineixen dos **escenaris** segons la disponibilitat de la variable `duration` (durada en segons de l'última trucada), que és coneguda únicament *a posteriori*. L'escenari POST_CALL la inclou i representa una situació d'anàlisi retrospectiva; l'escenari PRE_CALL l'exclou i és el cas realista de predicció *abans* de la trucada. Com que `duration` és amb diferència la variable més predictiva del model, la seva absència en PRE_CALL posa els sintetitzadors a prova en un escenari més exigent, sense el senyal dominant. La combinació de dos volums, dos escenaris i tres sintetitzadors dóna les **dotze configuracions** del banc de proves.

3.2 Pipeline experimental

El *pipeline* executa quatre fases sobre cada combinació:

1. **Preparació** (`prepare_data.py`): neteja, *split* estratificat 80/20 amb `seed=42` sobre `y` —proporció que garanteix un test de ∼8.000 registres, suficient per a estimacions d'AUC estables en un conjunt desequilibrat [11]—, `unknown` com a categoria i preservació del sentinella `pdays`.
2. **Entrenament del sintetitzador** (`train_copula.py`, `train_ctgan.py`, `train_tvae.py`) sobre `real_train`, amb $|synth| = |real_train|$.
3. **Avaluació d'utilitat** (`eval_tstr.py`): LR i RF entrenats sobre dades reals (TRTR) i sintètiques (TSTR), avaluats sobre `real_test`.
4. **Privacitat i qualitat** (`eval_privacy.py`, `eval_mia.py`, `eval_quality.py`, `eval_calibration.py`): DCR, MIA, SDMetrics i calibració; consolidat amb `consolidate_results.py`.

La tria de LR i RF com a models descendants respon a criteris de complementarietat diagnòstica: la LR és un model lineal inherentment calibrat (la funció sigmoide de sortida tendeix a produir probabilitats ben ajustades fins i tot sota *covariate shift* —el fenomen pel qual la distribució de les covariables d'entrenament i de test difereix, com passa quan s'entrena sobre dades sintètiques i es prova sobre dades reals— [11]) i permet detectar si el sintètic preserva les relacions lineals; el RF, no lineal i sensible al biaix de probabilitats, és un detector natural del Cas A de descalibració (§5.2). La seva comparació permet distingir artefactes d'avaluació de degradació estructural real. La mida del conjunt sintètic s'ha fixat en $|S| = |T_{\text{train}}|$, seguint la convenció habitual de la literatura [7, 9]: garanteix que el model descendent vegi el mateix volum de dades que en el baseline TRTR, fent directament comparable la diferència de rendiment. Els hiperparàmetres dels sintetitzadors corresponen als valors per defecte de SDV 1.18.0; el detall complet es recull a la taula 11 de l'apèndix.

3.3 Definició de les mètriques

A continuació es defineix cadascuna de les mètriques emprades, incloent-hi les fórmules on és rellevant per garantir la traçabilitat dels càlculs.

Utilitat (TSTR). L'eix d'utilitat s'avalua amb el protocol TSTR: s'entrena un model de classificació sobre dades sintètiques i s'avalua sobre el conjunt de test real (`real_test`). Com a sostre de referència s'usa el baseline TRTR, que representa el rendiment que s'obtindria entrenant directament amb dades reals. Les mètriques reportades són l'àrea sota la corba ROC (**AUC**), el F_1*-score* a llindar fix 0,5 i a llindar òptim. S'usen dues variants del F_1 perquè, en conjunts fortament desbalancejats, el llindar 0,5 no és l'òptim: un model ben ordenat pot tenir F_1@0,5 = 0 per mera descalibració de les probabilitats, i penalitzar-lo per això seria injust (vegeu §5.2).

Privacitat — DCR (*Distance to Closest Record*). La DCR quantifica la proximitat de cada registre sintètic al conjunt real d'entrenament. Per a un registre sintètic s_i, es defineix com:

$$\text{DCR}(s_i) \;=\; \min_{t_j \in train} d(s_i,\, t_j)$$

on $d(\cdot,\cdot)$ combina la distància euclidiana normalitzada per a variables numèriques i la distància de Hamming per a variables categòriques. Una DCR molt propera a zero indicaria una còpia quasi exacta d'un registre real, fet que seria problemàtic per a la privacitat. Per contextualitzar el valor, es compara la mediana de la DCR del conjunt sintètic (S) amb la mediana del *holdout* de test (on med denota la mediana de la distribució):

$$\text{DCR ratio} \;=\; \frac{\text{med}\big[\text{DCR}(S)\big]}{\text{med}\big[\text{DCR}(test)\big]} \qquad (1)$$

Un ratio $>$ 1 significa que el sintètic és, de mitjana, *més lluny* del train que la pròpia mostra de test, indicant

l'absència de còpies; un ratio < 1 indica una proximitat preocupant (§5.6). La intuïció és la següent: les dades de test mai han participat en l'entrenament del sintetitzador i representen la distància "natural" d'un observador extern. Usar el test com a referència permet respondre la pregunta rellevant: *el sintètic s'assembla als registres reals més del que ho fa qualsevol dada nova que no hagi tingut contacte amb el model?* Si la resposta és no (ratio > 1), el sintètic no aporta informació addicional sobre els individus d'entrenament.

Privacitat — MIA (*Membership Inference Attack*). L'atac d'inferència de pertinença [15] planteja la pregunta: *pot un atacant extern determinar si un registre concret ha format part del conjunt d'entrenament?* La implementació emprada és un atac basat en distàncies: per a cada registre sota atac, es calcula la seva distància mínima al conjunt sintètic com a indicador de si hi és "representat". L'AUC de la corba ROC d'aquest classificador binari indica la capacitat discriminativa de l'atac: un valor $\approx 0,5$ és equivalent a l'atzar i implica risc nul; un valor proper a 1 indicaria que l'atac identifica membres amb alta precisió. En termes pràctics: si l'AUC és 0,5, el conjunt sintètic és tan estadísticament "neutral" que fins i tot un algoritme dissenyat *específicament* per detectar membres d'entrenament no pot fer-ho millor que tirar una moneda a l'aire. Un AUC de 0,5 és, per tant, el millor resultat possible des del punt de vista de la privacitat.

Fidelitat estadística (SDMetrics). SDMetrics [4] quantifica la semblança estadística entre el conjunt sintètic i el real a través de dues dimensions: *Column Shapes*, que compara la distribució marginal de cada variable individualment (per exemple, si la distribució d'edats en el sintètic s'assembla a la real), i *Column Pair Trends*, que compara les correlacions bivariades entre parells de variables. L'*Overall* és la mitjana ponderada de les dues dimensions, en l'interval $[0, 1]$, on 1 indica semblança perfecta.

Calibració post-hoc. En un model de classificació, les probabilitats predites haurien de reflectir les freqüències observades: si el model assigna un 80 % de probabilitat a la classe positiva, hauria d'encertar en el 80 % dels casos. Si no és així, el model és *descalibrat*. La calibració de Platt (sigmoide) [12] i la calibració isotònica [13, 14] apliquen transformacions post-hoc sobre les probabilitats predites per apropar-les a les freqüències reals, sense tornar a entrenar el model.

3.4 Mostratge condicional, GPU i reproductibilitat

Per corregir el biaix de *class balance* de CTGAN del Seguiment 2 ($\Delta_{\text{rel}} = +81\%$), s'aplica el mostratge condicional (`sample_from_conditions`) sobre `y`, que força la proporció de classe sintètica a coincidir amb la real. Els ajustos de CTGAN i TVAE, costosos (el *fit* de CTGAN en CPU trigava $\sim$10,5 h), s'han externalitzat a Google Colab amb GPU (`device=cuda`); GaussianCopula s'executa localment en CPU en menys d'un minut. Tot és reproductible: `seed=42`, versions fixades (`sdv==1.18.0`, `sdmetrics==0.18.0`), resultats traçables a CSV amb `run_all_final.bat` i sintètics versionats a `data/processed/`. Cal notar que GaussianCopula és inherentment estable: el seu ajust (màxima versemblança de la còpula) és determinista i la llavor afecta únicament el *split* de dades, no l'estimació del model.

3.5 Procés de desenvolupament

El treball segueix tres fases: Seguiment 1 (baseline), Seguiment 2 (GaussianCopula, diagnòstic Cas A) i fase final (dotze combinacions, calibració, MIA).

4 RESULTATS

Aquesta secció presenta els resultats consolidats; la seva interpretació es desenvolupa a §5. El conjunt complet de marcadors per a les dotze combinacions (utilitat, privacitat i fidelitat) es recull a la taula 13 de l'apèndix A.5. A continuació es detallen els resultats per eix.

Utilitat. El baseline real→real (TRTR) es reprodueix dins la tolerància de 0,005 d'AUC. La taula 3 mostra la comparativa entre aquest baseline i el millor resultat TSTR per a cadascuna de les quatre configuracions (dataset × escenari): el millor sintetitzador reté entre el 92 % i el 98 % de l'AUC real, la qual cosa indica que les dades sintètiques preserven la major part de la capacitat predictiva. En règim TSTR, Logistic Regression manté entre el 92 % i el 97 % de l'AUC real en totes les configuracions. El Random Forest mostra un F_1@0,5 nul en les quatre configuracions de GaussianCopula i valors creixents en CTGAN i TVAE, un patró que la calibració i l'anàlisi AUC vs. F_1 expliquen a §5.

TAULA 3: TRTR (DADES REALS) VS. MILLOR TSTR (RF). "RET." = % DE L'AUC REAL RECUPERAT PEL MILLOR SINTETITZADOR.

Dataset	Esc.	TRTR AUC	TSTR AUC	Ret. (%)
add.	post	0,9491	0,8935 (tvae)	94,1
add.	pre	0,7844	0,7676 (ctgan)	97,9
full	post	0,9301	0,8798 (tvae)	94,6
full	pre	0,7960	0,7327 (tvae)	92,0

Privacitat. El MIA retorna un AUC pràcticament igual a 0,5 en les dotze combinacions. El *ratio* DCR ordena els sintetitzadors en un gradient clar: GaussianCopula manté el major marge (1,61–2,51), CTGAN un marge ajustat però > 1 (1,11–1,50) i TVAE cau per sota de la unitat (0,60–0,91). La fracció de distàncies zero és nul·la o pràcticament nul·la en tots els casos.

Fidelitat. El *SDMetrics Overall* es manté en una banda estreta (0,81–0,90), amb CTGAN lleugerament al capdavant; és l'eix que menys discrimina.

Comparativa LR vs. RF. La taula 4 recull l'AUC TSTR de Logistic Regression i Random Forest per a les dotze combinacions. Amb LR, GaussianCopula obté l'AUC més alt en el cas principal (0,9190), per davant de CTGAN

(0,8929) i TVAE (0,8755), un ordre que RF inverteix completament. Aquesta inversió es deu al Cas A de descalibració (§5.2): LR no pateix el desplaçament de probabilitats per sota de 0,5 que col·lapsa l'F_1 de RF.

TAULA 4: AUC TSTR (SYNTH→REAL) DE LR I RF PER A LES DOTZE COMBINACIONS. EN NEGRETA, EL MILLOR SINTETITZADOR PER EIX.

Dataset	Escenari	Synth	LR AUC	RF AUC
additional	post_call	copula	**0,9190**	0,8664
additional	post_call	ctgan	0,8929	0,8491
additional	post_call	tvae	0,8755	**0,8935**
additional	pre_call	copula	0,7596	0,6992
additional	pre_call	ctgan	**0,7898**	**0,7676**
additional	pre_call	tvae	0,7514	0,7096
full	post_call	copula	0,8562	0,7654
full	post_call	ctgan	0,8772	0,8596
full	post_call	tvae	**0,8884**	**0,8798**
full	pre_call	copula	0,7088	0,6193
full	pre_call	ctgan	0,7222	0,7034
full	pre_call	tvae	**0,7292**	**0,7327**

5 DISCUSSIÓ

Aquesta secció interpreta els resultats de §4 seguint l'ordre: *trade-off* global (§5.1), tres troballes clau (§§5.2–5.4), calibració, privacitat, comparativa LR/RF, robustesa i implicacions pràctiques.

Resum dels resultats a discutir. El resultat central és que els tres sintetitzadors es diferencien de manera consistent: GaussianCopula és el més conservador en privacitat però el que presenta una aparent pèrdua d'utilitat deguda a un artefacte d'avaluació (Cas A); TVAE és el més útil en termes absoluts però el que té el menor marge de privacitat i un biaix de proporció de classes en absència de variables molt predictives; CTGAN ocupa la posició intermèdia i és l'únic que permet controlar el *class balance* de manera directa. En les seccions que segueixen es desglossen les causes i conseqüències d'aquest patró, i es proposen les eines de mitigació corresponents.

5.1 Trade-off cross-synth: utilitat, privacitat i qualitat

D'acord amb la literatura sobre síntesi tabular [7, 20], s'esperava que els models neuronals (CTGAN, TVAE) superessin l'enfocament paramètric (GaussianCopula) en utilitat sobre conjunts amb dependències no lineals, i que GaussianCopula preservés millor la privacitat per la seva naturalesa paramètrica i la menor capacitat d'ajust a casos individuals. S'esperava també que la qualitat estadística (SDMetrics) fos alta en totes les arquitectures, donat que SDV l'optimitza explícitament. Els resultats obtinguts confirmen *parcialment* aquestes expectatives, però introdueixen matisos rellevants: la superioritat neuronal en utilitat s'inverteix quan el model descendent és LR (on copula resulta millor), degut a la descalibració; i la qualitat SDMetrics és efectivament alta però no discrimina entre arquitectures, cosa que la fa inadequada com a criteri de selecció únic.

L'avaluació creuada sobre les dotze combinacions mostra un *trade-off* força consistent entre utilitat, privacitat i qualitat, en què cap arquitectura domina les altres en els tres eixos alhora. La taula 5 il·lustra el patró sobre el cas principal i la figura 1 el resumeix per a totes les configuracions.

TAULA 5: EIXOS D'AVALUACIÓ PER SINTETITZADOR EN EL CAS PRINCIPAL (BANK-ADDITIONAL-FULL / POST_CALL).

synth	AUC RF	F_1@0,5	DCR	MIA	SDM
copula	0,8664	0,0000	2,4666	0,5044	0,8141
ctgan	0,8491	0,2230	1,3640	0,5035	0,8831
tvae	0,8935	0,4180	0,9121	0,5039	0,8767

La lectura conjunta revela una ordenació gairebé inversa entre privacitat i utilitat-a-llindar-fix. La proximitat al conjunt d'entrenament (mediana del *ratio* DCR) és més folgada en GaussianCopula (**1,61–2,51**), intermèdia en CTGAN (**1,11–1,50**) i inferior a la unitat en TVAE (**0,60–0,91**): el sintètic de TVAE acaba més a prop del train que la mostra de test. En sentit contrari, l'F_1 del RF al llindar 0,5 ordena els sintetitzadors al revés: GaussianCopula col·lapsa a **0,0000** en les quatre configuracions, TVAE assoleix el rang més alt (**0,18–0,44**) i CTGAN se situa al mig. La qualitat (SDMetrics *Overall*) es manté en una banda estreta (**0,81–0,90**) i no discrimina entre arquitectures: és en si mateixa una troballa rellevant. Tots tres sintetitzadors reprodueixen l'estructura estadística global —distribucions marginals i correlacions bivariades— de manera comparable; les diferències entre arquitectures resideixen en com gestionen les dependències fines sota condicions adverses (senyal feble, desbalanceig acusat) i en el grau de proximitat als registres d'entrenament, que és exactament el que capturen la utilitat TSTR i les mètriques de privacitat. Emprar SDMetrics com a criteri de selecció únic seria, per tant, insuficient.

Cal remarcar que l'**ordenació creixent de la privacitat** (copula $>$ ctgan $>$ tvae en el *ratio* DCR) és consistent amb les propietats estructurals de cada família: GaussianCopula genera mostres a partir de distribucions paramètriques que, per construcció, no poden memoritzar registres individuals; els models neuronals (CTGAN, TVAE) optimitzen una funció de pèrdua que minimitza la distància entre la distribució generada i la d'entrenament, cosa que els permet capturar dependències no lineals però, alhora, els predisposa a apropar-se als registres del train (vegeu §5.6).

5.2 Troballa 1 — Cas A: descalibració del RF (GaussianCopula)

El col·lapse aparent de l'F_1 al llindar 0,5 del RF amb GaussianCopula ($F_1 = 0,0000$) no és una pèrdua d'informació sinó una descalibració pura (Cas A): les probabilitats predites queden comprimides per sota del llindar (figura 2).

La clau per distingir descalibració de pèrdua estructural és comparar l'F_1 amb l'AUC. Tot i l'F_1 nul a 0,5, l'AUC es manté plenament utilitzable (**0,8664** en el cas principal). La distinció és fonamental: l'AUC mesura si el model *ordena* bé els registres (els positius haurien de tenir puntuació més alta que els negatius), mentre que l'F_1@0,5 mesura si a llindar fix el model *classifica* correctament. Imaginem un examinador que puntua tots els alumnes de manera correcta i consistent, però que ha fixat el llindar d'aprovat a 8 sobre 10 en un examen en què tothom ha tret entre 3 i 6: tothom suspèn ($F_1 = 0$), però el rànquing és perfecte (AUC = 1). En el Cas A, el sintetitzador comprimeix les probabilitats

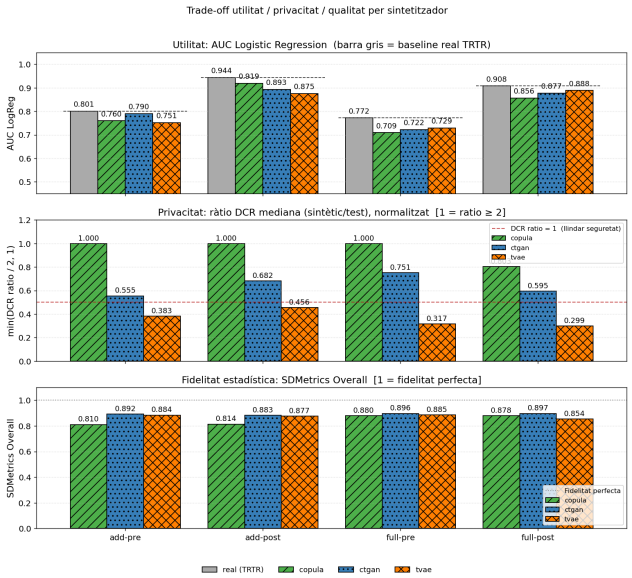

Fig. 1: Trade-off utilitat–privacitat–qualitat sobre les dotze combinacions, per sintetitzador (verd = GaussianCopula, blau = CTGAN, taronja = TVAE) i configuració (eix horitzontal). La barra grisa és el baseline real (TRTR). Panell superior: AUC de Logistic Regression en règim TSTR; la línia de guions marca el sostre real per configuració. Panell central: ràtio DCR normalitzat; la línia vermella indica el llindar de seguretat (ratio = 1). Panell inferior: SDMetrics *Overall*. Cap sintetitzador domina simultàniament els tres panells.

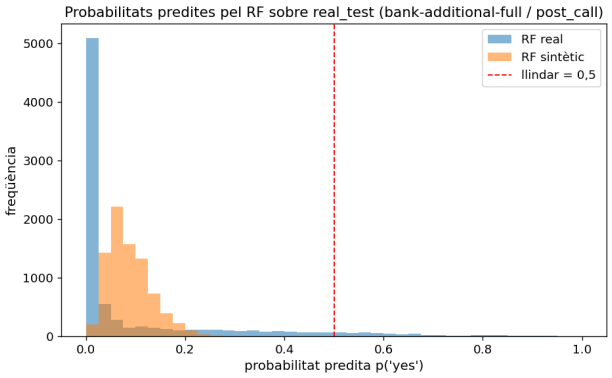

Fig. 2: Probabilitats predites pel RF (GaussianCopula, `bank-additional-full`/POST_CALL): tota la massa queda per sota de 0,5, fet que explica l'F_1 nul a llindar fix sense pèrdua de capacitat discriminativa (AUC 0,8664).

predites per sota del tall de 0,5; baixar el llindar (o recalibrar) recupera immediatament la utilitat. La conseqüència metodològica és que reportar només l'F_1 a 0,5 pot penalitzar incorrectament un sintetitzador que preserva la informació; la recuperació es desenvolupa a §5.5.

5.3 Troballa 2 — Biaix de class balance de TVAE en `pre_call`

En absència de `duration` (PRE_CALL), TVAE infrarepresenta de manera severa la classe positiva, un biaix simètric al biaix oposat de CTGAN sense mostratge condicional (Seguiment 2). La taula 6 en quantifica la magnitud.

TAULA 6: PREVALENÇA SINTÈTICA DE LA CLASSE POSITIVA (TVAE) VS. REAL.

dataset	scenari	pos synth	pos real	Δrel
additional	pre_call	0,0146	0,1127	-87%
additional	post_call	0,1200	0,1127	$\approx 0\%$
full	pre_call	0,0482	0,1170	-59%
full	post_call	0,0969	0,1170	-17%

El biaix és màxim en `additional`/PRE_CALL (pos_rate **0,0146** vs **0,1127**, Δrel -87%). El mecanisme subjacent és el següent: en l'escenari POST_CALL, la variable `duration` (durada de la trucada) actua com a ancoratge poderós de la classe positiva —els clients que subscriuen el producte tendeixen a tenir converses llargues— i el VAE aprèn a associar aquest senyal fort a la minoria. Quan `duration` s'elimina en PRE_CALL, l'ancoratge desapareix: sense un senyal que separi clarament els dos grups, el VAE aprèn la distribució global i en la generació tendeix a produir registres característics de la classe majoritària (88,3 % del conjunt). El resultat és que el sintètic és estadísticament correcte a nivell global —SDMetrics 0,88— però fracassa en reproduir la minoria. La simetria amb CTGAN (Δrel $\approx +81\%$ al Seg 2, on sobre-generava positius) mostra que cap model neuronal, entrenat sense restriccions, garanteix la proporció de classes del conjunt original.

CTGAN amb mostratge condicional com a solució d'enginyeria. El mostratge condicional força la proporció de classe a coincidir amb la real, fet que converteix CTGAN en l'única arquitectura que permet controlar el *class balance*: Δrel $\approx 0\%$ en les quatre configuracions, sense degradar la qualitat (SDMetrics **0,88–0,90**) ni travessar el llindar de privacitat (*ratio* DCR > 1).

5.4 Troballa 3 — Robustesa dissociada: AUC vs. F_1 a llindar fix

El biaix de prevalença de TVAE no col·lapsa la utilitat: desplaça el llindar òptim però preserva el rànquing, de manera que la degradació es concentra en l'F_1 a llindar fix i no en l'AUC (taula 7).

TAULA 7: TVAE: RETENCIÓ D'UTILITAT EN RÈGIM TSTR RESPECTE EL BASELINE TRTR, RANDOM FOREST.

dataset	scenari	AUC TSTR	AUC TRTR	F_1 TSTR	F_1 TRTR	ret. F_1
additional	pre_call	0,7096	0,7844	0,2304	0,3790	60,8%
additional	post_call	0,8935	0,9491	0,4180	0,5409	77,3%
full	pre_call	0,7327	0,7960	0,1841	0,3222	57,1%
full	post_call	0,8798	0,9301	0,4337	0,4619	93,9%

En els escenaris PRE_CALL la caiguda de l'AUC respecte del baseline és de només $\sim 9\%$ (0,7096 vs 0,7844 en `additional`; 0,7327 vs 0,7960 en `full`), mentre que la retenció de l'F_1 cau molt més (**57–61%** en PRE_CALL vs **77–94%** en POST_CALL). L'AUC és robust al biaix de prevalença; l'F_1 a 0,5 n'absorbeix tot el cost. El mecanisme és coherent amb la Troballa 1: amb pocs positius sintètics el RF prediu "yes" menys sovint (baixa el *recall* i l'F_1@0,5), però l'ordenació de probabilitats es manté.

5.5 Calibració post-hoc com a mitigació transversal

La calibració de Platt tanca el Cas A de GaussianCopula però resulta neutra o inclús contraproduent per a TVAE, perquè els dos problemes de fons no són el mateix: GaussianCopula pateix una *compressió* de probabilitats (totes les puntuacions queden per sota de 0,5, però l'ordre entre elles és correcte), mentre que TVAE pateix un *desplaçament de prevalença* (hi ha molt pocs positius sintètics, de manera que el model aprèn a donar puntuació baixa a la majoria de registres). La calibració de Platt corregeix la compressió perquè és una transformació monòtona que estira l'escala de probabilitats; però si el problema original és que *hi ha pocs positius* en les dades d'entrenament, recalibrar les puntuacions no afegeix positius: caldrà baixar el llindar de decisió per capturar-los. Ambdós símptomes donen F_1@0,5 baix, però els tractaments efectius són oposats (figura 3, taula 8).

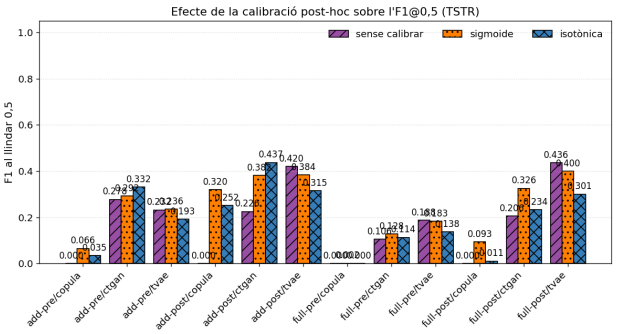

Fig. 3: Efecte de la calibració post-hoc sobre l'F_1@0,5 per sintetitzador i variant (*uncalib.*, sigmoide de Platt, isotònica), amb el llindar òptim (*f1@opt*) com a sostre de referència. La sigmoide recupera GaussianCopula de $F_1 = 0$ fins a 0,32–0,32; per a TVAE la palanca és l'ajust del llindar, no la recalibració.

TAULA 8: F_1@0,5 PER VARIANT DE CALIBRACIÓ. LA SIGMOIDE ÉS LA MILLOR PER A GAUSSIANCOPULA; PER A TVAE LA PALANCA ÉS L'AJUST DE LLINDAR (f_1@OPT).

synth	dataset/scenari	uncalib.	sigmoide	isotònica	f_1@opt
copula	additional/post	0,0000	0,3203	0,2524	0,4906
copula	full/post	0,0000	0,0935	0,0112	0,3612
copula	full/pre	0,0000	0,0019	0,0000	0,2583
tvae	additional/pre	0,2320	0,2361	0,1931	0,4234
tvae	additional/post	0,4204	0,3842	0,3148	0,5467

Nota: f_1@opt correspon a l'ajust de llindar òptim sobre les probabilitats sense recalibrar; les variants sigmoide i isotònica tenen llindars òptims propis amb valors molt propers ($\pm 0,003$ típic).

En el cas principal de GaussianCopula, la sigmoide recupera l'F_1 de **0,0000** a **0,3203**, i el llindar òptim fins a **0,4906** ($\sim$74% del rendiment real). Per a TVAE la calibració pot empitjorar lleugerament l'F_1@0,5 (`additional`/POST_CALL: 0,4204 $\rightarrow$ 0,3842 sigmoide), mentre que la millora prové del llindar (f_1@opt 0,5467). La calibració post-hoc no és universal sinó específica del mecanisme: corregeix la compressió de probabilitats (Copula) però no el biaix de prevalença (TVAE), on la palanca és el llindar.

5.6 Privacitat: concordança de DCR i MIA

Les dues mètriques coincideixen en el veredicte de fons — ni memorització ni senyal de pertinença— però són complementàries: el MIA descarta el risc de pertinença i la DCR matisa un gradient de proximitat que TVAE sacrifica (taula 9).

TAULA 9: MÈTRIQUES DE PRIVACITAT (RANG SOBRE LES DOTZE COMBINACIONS).

synth	MIA-AUC	DCR ratio
copula	0,4991–0,5044	1,61–2,51 (> 1)
ctgan	0,5020–0,5047	1,11–1,50 (> 1)
tvae	0,5039–0,5071	0,60–0,91 (< 1)

El MIA dóna un AUC pràcticament indistingible de l'atzar ($\approx \mathbf{0,50}$) en les dotze combinacions: un atacant no podria determinar millor que l'atzar si un registre va formar part del train. La DCR confirma l'absència de còpies (fracció de zeros nul·la) però afegeix un matís que el MIA no captura: la mediana del *ratio* DCR de TVAE és **inferior a la unitat** (0,60–0,91), és a dir, el sintètic és, de mitjana, més a prop del train que la mostra de test. No és memorització, però sí un grau de proximitat que convindria vigilar. Les dues mètriques no es contradiuen: el MIA acota el risc de pertinença i la DCR el de proximitat estructural.

TVAE (DCR ratio 0,60–0,91) no genera còpies punt a punt (fracció de zeros nul·la), però la literatura [20] situa el llindar de preocupació real en DCR ratio $< 0,5$; els valors de TVAE el superen amb marge ajustat. La distribució completa (figura 4, apèndix A.2) confirma l'absència de còpies però evidencia el desplaçament sistemàtic cap al train.

5.7 Comparativa LR vs. RF: impacte de la calibració sobre la tria

L'anàlisi de la taula 4 revela un matís rellevant: l'ordenació dels sintetitzadors per utilitat **depèn del model descendent emprat**. Amb Random Forest, TVAE lidera en el cas principal (AUC 0,8935) i GaussianCopula queda al final (0,8664). Amb Logistic Regression, l'ordre s'inverteix: Copula assoleix el primer lloc (AUC 0,9190) i TVAE passa a ser el darrer (0,8755) en el mateix escenari.

La causa és mecànica: la LR estima les probabilitats amb una funció sigmoide inherentment ben calibrada, de manera que no pateix el desplaçament per sota de 0,5 que causa el col·lapse del RF (Cas A). Per tant, si el model descendent és lineal, GaussianCopula és un candidat sòlid que combina la màxima privacitat amb una utilitat alta sense calibració addicional; CTGAN/TVAE resulten necessaris principalment quan s'utilitzen arbres o models no lineals.

5.8 Robustesa de les conclusions cross-dataset

El gradient de privacitat (copula $>$ ctgan $>$ tvae) i el patró de col·lapse del RF amb GaussianCopula es reprodueixen íntegrament en tots dos volums. En PRE_CALL, els indicadors macroeconòmics de `bank-additional-full` (`euribor3m`, `nr.employed`) aporten un senyal alternatiu que eleva l'AUC TSTR fins a 9 punts per sobre de `bank-full`; en POST_CALL, on `duration` domina, les

diferències inter-volum queden per sota de 3 punts. La consistència entre volums confirma que els resultats no depenen del conjunt concret sinó de les propietats estructurals de les arquitectures estudiades.

5.9 Implicacions per a la pràctica professional

La taula 10 sintetitza la recomanació contextual; aquí es concreten els criteris de tria. **Privacitat màxima:** GaussianCopula (<1 min CPU, DCR ratio 1,6–2,5) és la millor opció; si el model descendent és basat en arbres, aplicar calibració de Platt corregeix el Cas A fins al 74 % del rendiment real. **Ús general:** CTGAN amb mostratge condicional (GPU, ∼12 min, taula 12 de l'apèndix A.4) és l'única arquitectura que garanteix $\Delta\text{rel} \approx 0\%$ mantenint DCR ratio > 1; recomanat per defecte en conjunts desequilibrats. **Utilitat bruta post-trucada:** TVAE (GPU, ∼5 min) ofereix la millor retenció d'AUC ($\leq$94,6 % RF en POST_CALL), però el DCR ratio < 1 requereix supervisió explícita del risc. Les tres eines (SDV, scikit-learn, SDMetrics) són de codi obert; el *pipeline* s'automatitza amb `run_all_final.bat` i és reproduïble sense infraestructura especialitzada.

5.10 Síntesi: resposta a la hipòtesi ≪cap sintetitzador domina≫

Els resultats confirmen la hipòtesi: cap sintetitzador domina simultàniament en els tres eixos, de manera que la tria és una decisió sobre quin punt del compromís es prioritza (taula 10).

TAULA 10: RESUM DELS TRES EIXOS PER SINTETITZADOR.

Eix	Copula	CTGAN+cond	TVAE
Utilitat (F_1@0,5)	nul·la*	intermèdia	**millor**
Privacitat (DCR)	**millor**	ok	fluixa
Class balance	fidel	**fidel (forçat)**	esbiaixat (pre)
Qualitat (SDM)	0,81	**0,88–0,90**	0,85–0,88
Cost	<1 min CPU	GPU	GPU
Recomanat per a	Privacitat màx.	**Ús general**	Utilitat màx.

* recuperable amb calibració/llindar.

5.11 Limitacions de l'avaluació

Una única llavor, sense intervals de confiança. Tots els resultats provenen d'una sola execució (`seed=42`); les diferències petites s'han d'interpretar com a tendència, no com a resultats estadísticament significatius. El treball és, en aquest sentit, un **primer prototip o aproximació exploratòria**: l'objectiu principal és establir el marc d'avaluació i identificar els patrons qualitatius del *trade-off*, no quantificar-los amb precisió estadística. Els gradients principals (rang de privacitat DCR, dissociació AUC vs. F_1) són prou amplis per ser robustos a una sola llavor; repetir amb múltiples llavors per obtenir intervals de confiança és la primera línia de treball futur (§6).

Llindar òptim calculat sobre `real_test`. El llindar òptim que recupera l'F_1 de GaussianCopula es cerca sobre el test. Aquest ús és **diagnòstic**, no **operatiu**: demostra que el rànquing és recuperable i que el col·lapse és un Cas A, no reporta un rendiment desplegable. En producció el llindar es fixaria sobre validació. Cap conclusió depèn de tractar l'F_1 òptim com a rendiment operatiu.

Reproductibilitat: fits GPU. CTGAN i TVAE es van ajustar en GPU (Colab, `device=cuda`); el determinisme exacte no és garantit amb `cuda`. GaussianCopula és determinista en CPU (§3). Els sintètics queden versionats com a evidència.

Hiperparàmetres per defecte sense ajust. La comparativa empra els valors per defecte de SDV 1.18.0 sense *tuning*; no es pot descartar que afavoreixin alguna arquitectura. Per a GaussianCopula l'ajust és tancat (màxima versemblança), però per a CTGAN i TVAE els valors per defecte d'epochs i arquitectura constitueixen una elecció implícita que limita la interpretació de les diferències de rendiment de segon ordre.

6 CONCLUSIONS

El treball ha assolit els cinc objectius. **(O1)** S'ha construït un *pipeline* reproductible PREPARE→TRAIN→EVAL, amb llavor fixa i resultats traçables. **(O2)** S'ha sintetitzat i avaluat el baseline GaussianCopula sobre les quatre configuracions. **(O3)** El baseline real→real s'ha validat dins la tolerància de 0,005 d'AUC. **(O4)** La comparativa s'ha completat amb CTGAN i TVAE fins a les dotze combinacions, amb dues mètriques de privacitat. **(O5)** S'ha discutit el *trade-off* entre els tres eixos, que constitueix el principal resultat del treball.

Resposta a la hipòtesi. Cap sintetitzador domina simultàniament. GaussianCopula maximitza la privacitat (*ratio* DCR 1,6–2,5) a costa d'una utilitat-RF nul·la a llindar fix —recuperable fins a ∼74% del rendiment real; TVAE inverteix el compromís (millor F_1 però *ratio* DCR < 1 i biaix de prevalença en PRE_CALL); CTGAN amb mostratge condicional ocupa el punt d'equilibri i és l'únic que permet controlar el *class balance*. La tria és una decisió sobre el punt desitjat del compromís utilitat–privacitat, no sobre la qualitat absoluta, que no discrimina (SDMetrics 0,81–0,90).

Contribució metodològica. (i) Avaluar la utilitat amb el llindar òptim i no només a 0,5, per no penalitzar models ben ordenats però descalibrats; (ii) reportar la privacitat amb almenys dues mètriques complementàries, atès que MIA i DCR acoten riscos ortogonals (pertinença vs. proximitat estructural); i (iii) documentar que el mostratge condicional resol el biaix de *class balance* com a decisió d'enginyeria independent de la qualitat intrínseca del sintetitzador, un aspecte ignorat en la majoria de comparatives públiques. Les tres contribucions tenen valor transversal: s'apliquen a qualsevol conjunt tabular desequilibrat i a qualsevol sintetitzador que produeixi probabilitats de sortida, no únicament als tres estudiats en aquest treball. El Cas A i el biaix de prevalença apareixerien en qualsevol estudi similar sense una anàlisi AUC vs. F_1 explícita. En perspectiva, les tres contribucions reforcen la hipòtesi central: la necessitat del diagnòstic per Cas A (A1), la complementarietat

MIA+DCR per acotare riscos ortogonals (A2) i el mostratge condicional com a decisió d'enginyeria independent de la qualitat intrínseca (A3) emergeixen directament de la impossibilitat de cap sintetitzador de dominar simultàniament els tres eixos.

Treball futur. Les línies de continuació prioritàries són: (i) repetir els experiments amb múltiples llavors per obtenir intervals de confiança i confirmar la significació estadística; (ii) ampliar a dominis variats (salut, finances) per verificar la generalitzabilitat del gradient de privacitat; (iii) explorar el *tuning* d'hiperparàmetres de CTGAN i TVAE per reduir la bretxa de privacitat sense mostratge condicional explícit; i (iv) incorporar privacitat diferencial formal [16] o arquitectures de difusió tabulars [18], una línia emergent que ha superat GAN i VAE en *benchmarks* recents.

DECLARACIÓ D'ÚS D'EINES D'INTEL·LIGÈNCIA ARTIFICIAL

En la redacció d'aquest informe s'han emprat assistents d'intel·ligència artificial generativa com a suport per a la revisió d'estil, la generació d'estructura de seccions, la consolidació de xifres dels CSVs en taules LaTeX i la verificació creuada de coherència entre text i resultats. Tots els resultats numèrics, taules i conclusions provenen de l'execució real dels *scripts* de `src/` sobre les dades reals; cap xifra ha estat generada o estimada per l'eina. La interpretació metodològica, el disseny dels experiments i les decisions sobre l'enfocament són responsabilitat de l'autor.

REFERÈNCIES

[1] S. Moro, P. Cortez, P. Rita. *A data-driven approach to predict the success of bank telemarketing.* Decision Support Systems, 62:22–31, 2014.

[2] N. Patki, R. Wedge, K. Veeramachaneni. *The Synthetic Data Vault.* A *IEEE DSAA*, pàgines 399–410, 2016. doi:10.1109/DSAA.2016.49.

[3] The Synthetic Data Vault project. *SDV documentation.* https://docs.sdv.dev/sdv/, consultat el 22 de maig de 2026.

[4] The Synthetic Data Vault project. *SDMetrics documentation.* https://docs.sdv.dev/sdmetrics/, consultat el 22 de maig de 2026.

[5] A. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Publ. Inst. Statist. Univ. Paris, 8:229–231, 1959.

[6] I. Goodfellow et al. *Generative Adversarial Networks.* A *NeurIPS*, vol. 27, pàgines 2672–2680, 2014.

[7] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni. *Modeling Tabular Data using Conditional GAN.* A *NeurIPS*, vol. 32, pàgines 7335–7345, 2019. arXiv:1907.00503. (Introdueix CTGAN i TVAE.)

[8] D. P. Kingma, M. Welling. *Auto-Encoding Variational Bayes.* A *ICLR*, 2014. arXiv:1312.6114.

[9] C. Esteban, S.L. Hyland, G. Rätsch. *Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs.* A *NIPS Workshop on Machine Learning for Health*, 2017. arXiv:1706.02633. (Protocol TSTR.)

[10] L. Breiman. *Random Forests.* Machine Learning, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.

[11] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python.* JMLR, 12:2825–2830, 2011.

[12] J. Platt. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.* A *Advances in Large Margin Classifiers*, MIT Press, pàgines 61–74, 1999.

[13] B. Zadrozny, C. Elkan. *Transforming Classifier Scores into Accurate Multiclass Probability Estimates.* A *ACM SIGKDD*, pàgines 694–699, 2002.

[14] A. Niculescu-Mizil, R. Caruana. *Predicting Good Probabilities with Supervised Learning.* A *ICML*, pàgines 625–632, 2005.

[15] R. Shokri, M. Stronati, C. Song, V. Shmatikov. *Membership Inference Attacks Against Machine Learning Models.* A *IEEE S&P*, pàgines 3–18, 2017. doi:10.1109/SP.2017.41.

[16] C. Dwork. *Differential Privacy.* A *ICALP*, pàgines 1–12, 2006.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. *SMOTE: Synthetic Minority Over-sampling Technique.* JAIR, 16:321–357, 2002.

[18] A. Kotelnikov, D. Baranchuk, I. Rubachev, A. Babenko. *TabDDPM: Modelling Tabular Data with Diffusion Models.* A *ICML*, pàgines 17564–17579, 2023. arXiv:2209.15421.

[19] Z. Zhao, A. Kunar, R. Birke, L. Y. Chen. *CTAB-GAN: Effective Table Data Synthesizing.* A *ACML*, pàgines 97–112, 2021. arXiv:2102.08369.

[20] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, P. Cohen, A. Weller. *Synthetic Data — what, why and how?* arXiv:2205.03257, 2022. Informe encarregat per The Royal Society.

[21] A. G. Solatorio, O. Dupriez. *REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers.* arXiv:2302.02041, 2023.

APÈNDIX

A.1 Diagnòstic del col·lapse del Random Forest (Cas A)

El Seguiment 2 va contrastar tres hipòtesis ortogonals per al $F_1 = 0$ del RF amb GaussianCopula: descalibració (Cas A), esbiaix de mostratge (Cas B) i divergència de *feature importances* (Cas C). El veredicte va ser Cas A pur en POST_CALL i mixt A+C en PRE_CALL. La figura 2 (§5.2) n'és l'evidència visual principal.

A.2 Sèries de privacitat per configuració (DCR)

La figura 4 mostra la distribució de la DCR del conjunt sintètic i del *holdout* de test per als tres sintetitzadors en el cas principal (`bank-additional-full` / POST_CALL), i il·lustra el gradient de proximitat descrit a §5.6: GaussianCopula manté el sintètic més lluny del train, CTGAN un marge ajustat i TVAE el situa més a prop.

A.3 Hiperparàmetres dels sintetitzadors

La taula 11 recull els hiperparàmetres principals de CTGAN i TVAE tal com s'han emprat en aquest treball, corresponents als valors per defecte de `sdv==1.18.0`. GaussianCopula no té hiperparàmetres configurables a part del tipus de marginals, que SDV infereix automàticament.

TAULA 11: HIPERPARÀMETRES PRINCIPALS DELS SINTETITZADORS NEURALS (SDV 1.18.0). "—" INDICA QUE EL PARÀMETRE NO ÉS APLICABLE A L'ARQUITECTURA.

Paràmetre	**CTGAN**	**TVAE**
`epochs`	300	300
`batch_size`	500	500
`generator_dim`	(256, 256)	—
`discriminator_dim`	(256, 256)	—
`pac`	10	—
`compress_dims`	—	(128, 128)
`decompress_dims`	—	(128, 128)
`l2scale`	—	1e-5
`loss_factor`	—	2
`device`	cuda	cuda

A.4 Cost computacional (`fit_metadata`)

La taula 12 recull el temps d'ajust de cada sintetitzador, traçat a `fit_metadata.csv`. GaussianCopula s'entrena en CPU en menys d'un minut, mentre que CTGAN i TVAE requereixen GPU (Google Colab, `device=cuda`); el cost de CTGAN és notablement superior pel mostratge condicional.

TAULA 12: TEMPS D'AJUST (*fit*) PER SINTETITZADOR I CONFIGURACIÓ.

Synth	Configuració	fit (min)	device
copula	totes (4)	< 1	cpu
ctgan	additional/post	12,70	cuda
ctgan	additional/pre	12,39	cuda
ctgan	full/post	12,39	cuda
ctgan	full/pre	11,95	cuda
tvae	additional/post	5,64	cuda
tvae	additional/pre	5,24	cuda
tvae	full/post	5,36	cuda
tvae	full/pre	5,04	cuda

A.5 Marcadors consolidats de les dotze combinacions

La taula 13 recull, per a cada combinació, els marcadors dels tres eixos: utilitat (AUC i F_1 del RF en règim TSTR amb i sense calibració), privacitat (mediana del *ratio* DCR i AUC del MIA) i fidelitat estadística (SDMetrics *Overall*).

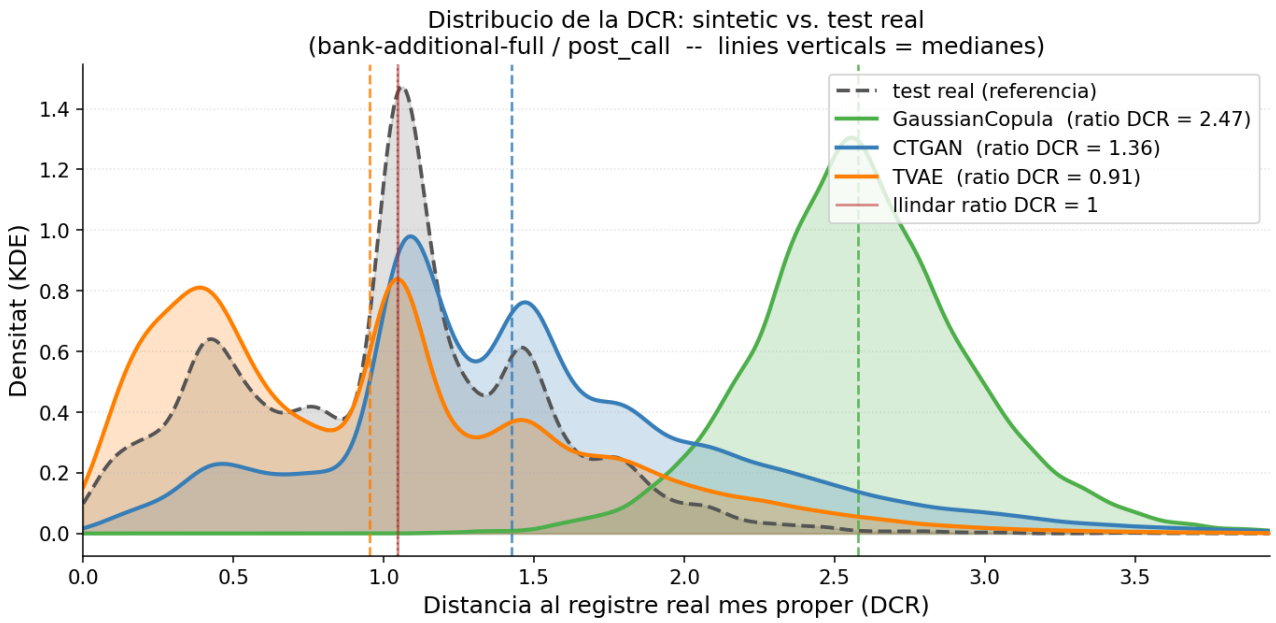

Fig. 4: Distribució de la *Distance to Closest Record* (DCR) en el cas principal (`bank-additional-full`/POST_CALL). Les corbes KDE dels tres sintetitzadors (verd = GaussianCopula, blau = CTGAN, taronja = TVAE) es superposen sobre la distribució de referència del *holdout* de test (gris, línia discontínua). Les línies verticals discontínues indiquen la mediana de cada distribució sintètica; la línia vermella assenyala la mediana del test (llindar ratio DCR = 1). GaussianCopula (ratio 2,47) presenta el sintètic desplaçat clarament a la dreta (major distància, major privacitat), CTGAN (ratio 1,36) en posició intermèdia, i TVAE (ratio 0,91) per sota del llindar.

TAULA 13: MARCADORS CONSOLIDATS DE LES DOTZE COMBINACIONS (RANDOM FOREST, TSTR SYNTH→REAL). *Calib.* = MILLOR F_1@0,5 AMB CALIBRACIÓ POST-HOC.

Dataset	Escenari	Synth	AUC RF	F_1@0,5	F_1 calib.	DCR ratio	MIA-AUC	SDM overall
additional	post_call	copula	0,8664	0,0000	0,3203	2,4666	0,5044	0,8141
additional	post_call	ctgan	0,8491	0,2230	0,4373	1,3640	0,5035	0,8831
additional	post_call	tvae	0,8935	0,4180	0,4204	0,9121	0,5039	0,8767
additional	pre_call	copula	0,6992	0,0000	0,0658	2,5128	0,4991	0,8101
additional	pre_call	ctgan	0,7676	0,2768	0,3315	1,1095	0,5047	0,8922
additional	pre_call	tvae	0,7096	0,2304	0,2361	0,7650	0,5071	0,8837
full	post_call	copula	0,7654	0,0000	0,0935	1,6055	0,4997	0,8781
full	post_call	ctgan	0,8596	0,2046	0,3255	1,1890	0,5020	0,8968
full	post_call	tvae	0,8798	0,4337	0,4357	0,5980	0,5047	0,8539
full	pre_call	copula	0,6193	0,0000	0,0019	2,0034	0,5039	0,8800
full	pre_call	ctgan	0,7034	0,1013	0,1284	1,5012	0,5033	0,8962
full	pre_call	tvae	0,7327	0,1841	0,1881	0,6348	0,5045	0,8850

A.6 Planificació temporal del projecte

La figura 5 recull la planificació setmanal del projecte, distribuïda en 21 setmanes (del 09/02 al 05/07 de 2026), estructurada en quatre fites oficials: l'**Informe Inicial** (09/03), el **Seguiment 1** (19/04), el **Seguiment 2** (24/05) i el **Lliurament Final** (28/06). El seguiment ha estat bo; l'única desviació va ser el diagnòstic no previst del Cas A (§5.2), que va consumir ∼2 setmanes addicionals al segon bloc (setmanes 6–14). Les contribucions metodològiques A1–A3 es van incorporar entre el Seguiment 1 i el Seguiment 2 sense modificar la data de lliurament.

Tasques	Inici (setm.)	Fi (setm.)	Notes	09–15 /02	16–22 /02	23–01 /03	02–08 /03	09–15 /03	16–22 /03	23–29 /03	30–05 /04	06–12 /04	13–19 /04	20–26 /04	27–03 /05	04–10 /05	11–17 /05	18–24 /05	25–31 /05	01–07 /06	08–14 /06	15–21 /06	22–28 /06	29–05 /07
Definició del tema i abast del TFG	1	2	Concretar problema, objectius i criteris d'avaluació																					
Revisió bibliogràfica i estat de l'art	1	4	SDV, dades sintètiques tabulars, mètriques (qualitat, utilitat, privadesa)																					
Planificació i estructura de l'informe inicial	2	4	Estructura Overleaf, apartats i cronograma																					
Selecció dataset (pendent de validació amb el tutor)	3	5	Tancament del dataset i variable objectiu																					
Disseny metodologia i pla experimental	3	5	Pipeline SDV (fit→sample→evaluate), models i mètriques																					
Lliurament: Informe inicial (09/03)	**5**	**5**	**Fita oficial**																					
Preprocessat + metadata final	5	7	Nuls, tipologies, codificació i validació																					
Baseline ML real→real	6	7	Models senzills i mètriques de referència																					
Implementació pipeline reproduïble (seeds, configs)	7	9	Scripts/notebooks i registre d'experiments																					
Entrenament sintetitzadors i generació (Copula/CTGAN/TVAE)	8	11	Versions sintètiques i ajustos bàsics																					
Avaluació de qualitat (distribucions, dependències)	9	10	Taules i figures de qualitat																					
Preparació seguiment 1	9	9	Resum de progrés i decisions																					
Lliurament: Informe seguiment 1 (19/04)	**10**	**10**	**Fita oficial**																					
Avaluació utilitat downstream (TSTR)	11	14	Train sintètic → test real (AUC/F1/Accuracy)																					
Avaluació privadesa/risc (proximitat/sobreajust)	12	15	Indicadors de memorització i risc																					
Discussió trade-off utilitat vs privadesa	15	16	Selecció millor model i justificació																					
Preparació seguiment 2	14	14	Consolidació de resultats																					
Lliurament: Informe seguiment 2 (24/05)	**15**	**15**	**Fita oficial**																					
Redacció informe final (format article) + figures	16	19	Metodologia, resultats, discussió i conclusions																					
Dossier del TFG (codi, manuals, annexos)	16	20	Compendi tècnic i reproductibilitat																					
Revisions finals amb el tutor	19	20	Correccions i poliment																					
Lliurament: Dossier + Informe final (28/06)	**20**	**20**	**Fita oficial**																					
Preparació presentació i assaig	20	21	Diapositives, guió i preguntes																					
Defensa al tribunal (30/06)	**21**	**21**	**Presentació i debat públic**																					

Fig. 5: Diagrama de Gantt del TFG (curs 2025/26). Les files grogues corresponen a les fites oficials; les verdes als lliuraments intermedis.