UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

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

Jaime Cabrera, Alejandro. *Evaluación de modelos generativos para datos sintéticos tabulares*. 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/337345

under the terms of the CC BY-SA license.

# Evaluación de modelos generativos para datos sintéticos tabulares

Alejandro José Jaime Cabrera

**Resumen –** La generación de datos sintéticos puede facilitar el análisis, la experimentación y la compartición de datos cuando el acceso directo a registros reales está limitado por privacidad, confidencialidad o disponibilidad. Este trabajo evalúa modelos generativos para datos tabulares bajo un protocolo común que combina utilidad predictiva, fidelidad estadística y señales empíricas de privacidad. Se comparan cinco datasets públicos y cuatro familias de generadores: Gaussian Copula, CTGAN, TVAE y TabDDPM. Los resultados muestran que no existe un generador universalmente superior: TVAE preserva mejor la utilidad media, Gaussian Copula y TVAE obtienen buena fidelidad global, CTGAN suele evitar duplicación literal y TabDDPM mantiene mayor separación respecto al conjunto de entrenamiento en varios datasets pero con resultados más irregulares. El caso Mushroom muestra que una utilidad muy alta puede coexistir con duplicación elevada, por lo que la generación de datos sintéticos no debe confundirse con anonimización formal.

**Palabras clave –** datos sintéticos, datos tabulares, CTGAN, TVAE, TabDDPM, privacidad empírica, TSTR, SDMetrics.

**Abstract –** Synthetic data generation can support analysis, experimentation and data sharing when direct access to real records is constrained by privacy, confidentiality or availability. This project evaluates tabular generative models through a common protocol combining predictive utility, statistical fidelity and empirical privacy signals. Five public datasets and four generator families are compared: Gaussian Copula, CTGAN, TVAE and TabDDPM. Results show that no generator is universally best: TVAE achieves the highest average utility, while Gaussian Copula and TVAE provide high global fidelity, CTGAN tends to avoid literal duplication, and TabDDPM often remains farther from the training data but behaves less consistently. The Mushroom case shows that high predictive utility can coexist with substantial duplication, so synthetic data should not be interpreted as formal anonymization.

**Keywords –** synthetic data, tabular data, CTGAN, TVAE, TabDDPM, empirical privacy, TSTR, SDMetrics.

## 1 Introducción y objetivos

### 1.1 Motivación

El uso de datos es central para entrenar, validar y comparar modelos de aprendizaje automático. Sin embargo, en muchos contextos reales su uso queda condicionado por privacidad, regulación, confidencialidad o limitaciones prácticas de acceso. La generación de datos sintéticos se plantea como una alternativa: crear registros artificiales que preserven propiedades relevantes del conjunto original y permitan prototipado, análisis o validación reduciendo la exposición directa de registros reales.

Sin embargo, un conjunto de datos sintéticos no debe interpretarse automáticamente como anónimo. Por ello, el proyecto no se limita a medir parecido estadístico, sino que estudia conjuntamente fidelidad, utilidad predictiva y señales empíricas de privacidad.

- Trabajo tutorizado por Guillermo Navarro (UAB).
- Grado en Ingeniería de Datos. Curso 2025/26.

### 1.2 Objetivos

El objetivo general del TFG es analizar el estado actual de la generación de datos sintéticos con modelos de aprendizaje profundo y evaluar experimentalmente su aplicación a datos tabulares. A partir de este objetivo, el trabajo se organiza en los siguientes objetivos específicos, ordenados por prioridad y dependencia metodológica.

**Objetivos principales**

1. Revisar el estado del arte y las herramientas disponibles para la generación de datos sintéticos tabulares.
2. Delimitar un benchmark tabular reproducible mediante la selección y caracterización de un conjunto reducido de datasets públicos y una variedad representativa de retos tabulares.
3. Seleccionar al menos dos generadores adecuados para el estudio experimental, priorizando su relevancia en la literatura y su viabilidad dentro del alcance del TFG.
4. Establecer un conjunto de métricas y un protocolo experimental apropiados para evaluar la calidad, la utili-

dad y los aspectos de privacidad de los datos sintéticos generados.

5. Ejecutar el entrenamiento, la generación y la evaluación de los modelos seleccionados conforme al protocolo definido.
6. Analizar los resultados obtenidos e identificar los principales trade-offs entre calidad, utilidad y privacidad.
7. Documentar de forma organizada el proceso, el código, los resultados y las conclusiones finales para asegurar la reproducibilidad del trabajo y la correcta elaboración del dossier.

**Objetivos secundarios**

8. Desarrollar un módulo de perfilado de datos para caracterizar los datasets y justificar los distintos retos tabulares considerados.
9. Ampliar el estudio con modelos adicionales o incorporar métricas complementarias que aporten valor comparativo al análisis.
10. Incorporar un threat model básico para interpretar con más contexto los resultados de privacidad.

Estos objetivos secundarios se integran en el desarrollo del benchmark mediante el perfilado de datasets, la ampliación de modelos y la interpretación de privacidad mediante un threat model empírico.

### 1.3 Organización y evolución del trabajo

El trabajo se organizó inicialmente como una evaluación incremental. La prioridad fue construir una infraestructura mínima viable, validar el flujo experimental con modelos y datasets base, y ampliar el alcance solo cuando los resultados fueran suficientemente reproducibles. Esta forma de trabajo permitió mantener el plan inicial como guía, pero adaptarla de forma controlada a medida que el estudio experimental avanzaba.

En conjunto, la planificación no se sustituyó, sino que se amplió por etapas. Primero se validó una infraestructura mínima con dos modelos y datasets base; después se reforzó la evaluación con perfilado y varias semillas; finalmente se incorporaron modelos adicionales y análisis complementarios. La correspondencia resumida entre el plan inicial y la evolución real del trabajo se recoge en la Tabla A.1 del apéndice.

### 1.4 Alcance y pregunta de investigación

El alcance se limita a datos tabulares *single-table*. Cada fila representa una entidad y cada columna un atributo numérico o categórico. Esta decisión mantiene el proyecto en un escenario realista, pero evita la complejidad adicional de bases de datos relacionales, series temporales o grafos. La pregunta de investigación final es: ¿Cómo se comparan modelos estadísticos clásicos, GANs, VAEs y modelos de difusión para datos tabulares en términos de utilidad, fidelidad y privacidad empírica?

## 2 CONTEXTO Y ESTADO DEL ARTE

Los datos sintéticos pueden generarse con enfoques estadísticos clásicos o con modelos generativos profundos. En el contexto tabular, la dificultad proviene de la heterogeneidad: variables continuas y categóricas, cardinalidades distintas, valores faltantes, desbalance de clases, distribuciones sesgadas, outliers y dependencias entre columnas. A diferencia de imagen o texto, una fila tabular no suele tener una estructura espacial o secuencial natural, por lo que el preprocesado y la representación de variables son parte esencial del problema.

En este trabajo, un modelo generativo se entiende como un método que aprende patrones de un conjunto de datos real con el objetivo de producir nuevas muestras artificiales. A diferencia de un modelo predictivo, que recibe una fila y devuelve una etiqueta o una probabilidad, un generador intenta aproximar la distribución conjunta de las variables: qué valores aparecen, con qué frecuencia y cómo se relacionan entre sí. En datos tabulares, esto implica capturar tanto distribuciones individuales de columnas como dependencias entre variables numéricas y categóricas.

El benchmark compara cuatro familias. **Gaussian Copula** actúa como baseline estadístico. La idea general de una cópula es separar la distribución individual de cada columna de la estructura de dependencia entre columnas [9]. En términos simples, el modelo intenta reproducir cómo se comporta cada variable por separado y cómo tienden a variar unas con otras. Aunque no es un modelo de deep learning, resulta útil como referencia sencilla, interpretable y competitiva.

**CTGAN** representa la familia de GANs adaptadas a datos tabulares. Una GAN se basa en una competición entre dos redes neuronales: un generador, que produce muestras artificiales, y un discriminador, que intenta distinguir entre datos reales y sintéticos. Durante el entrenamiento, el generador mejora intentando engañar al discriminador, mientras el discriminador mejora detectando muestras generadas. CTGAN adapta este esquema mediante entrenamiento condicional y transformaciones específicas para columnas continuas y categóricas [1].

**TVAE** aplica el marco de autoencoders variacionales a datos tabulares. Un VAE utiliza una arquitectura de codificador-decodificador: el codificador transforma cada fila real en una representación comprimida y probabilística, llamada espacio latente, y el decodificador aprende a reconstruir datos a partir de esa representación. Para generar nuevas muestras, el modelo toma puntos del espacio latente y los decodifica como filas plausibles. TVAE adapta este enfoque a variables tabulares numéricas y categóricas [2, 1].

Los modelos de difusión siguen una lógica distinta. Durante el entrenamiento se añade ruido progresivamente a los datos hasta degradarlos, y el modelo aprende el proceso inverso: eliminar ruido paso a paso para reconstruir muestras coherentes. En generación, se parte de ruido y se aplica ese proceso aprendido de denoising hasta obtener una fila de datos sintéticos. **TabDDPM** adapta esta familia de modelos a datos tabulares [3, 4].

La evaluación de datos sintéticos no puede reducirse a una única métrica. Una muestra puede ser fiel estadísticamente pero poco útil para entrenar modelos, o puede

ofrecer alta utilidad reproduciendo demasiado de cerca los registros de entrenamiento. Por ello, el trabajo adopta una lectura multidimensional basada en literatura reciente sobre datos sintéticos [5, 6]. En la parte práctica se usa SDV (Synthetic Data Vault) como ecosistema de herramientas para modelar y generar datos sintéticos tabulares, y SDMetrics como librería de evaluación para comparar datos reales y sintéticos mediante métricas y reportes específicos [7, 8].

## 3 METODOLOGÍA

La metodología se diseñó como un benchmark experimental organizado en cinco fases: selección y perfilado de datasets, configuración de generadores, ejecución multiseed, evaluación con métricas complementarias y análisis agregado. La implementación del benchmark se desarrolló en Python, apoyándose en configuraciones por dataset, scripts reproducibles y salidas agregadas en formato tabular.

### 3.1 Selección y perfilado de datasets

Antes de fijar el benchmark final se perfiló una bolsa inicial de 21 datasets públicos de clasificación disponibles en el UCI Machine Learning Repository [16]. El perfilado midió tamaño, tipos de variables, cardinalidad, missing, desbalance, duplicados, asimetría distribucional y atípicos. A partir de esos resultados se seleccionaron cinco datasets, no por cantidad, sino para cubrir extremos complementarios del espacio tabular [11, 12, 13, 14, 15]. La Figura 1 muestra el perfil normalizado usado para justificar la selección.

La lectura por dataset se resume así: **Adult** combina tipos de variables y cardinalidad; **Bank Marketing** concentra desbalance y asimetría distribucional; **Default Credit** tensiona el tratamiento de variables numéricas con atípicos; **Dry Bean** aporta un escenario numérico multiclase con desbalance y duplicados; y **Mushroom** actúa como caso categórico dominado por modos categóricos.

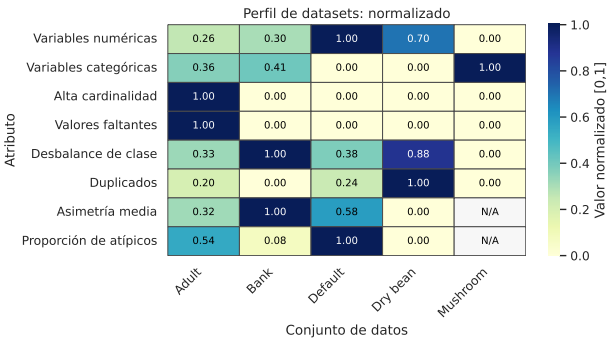


**Fig. 1:** Perfil normalizado de los datasets según retos tabulares. Los valores se normalizan a [0,1] dentro de cada atributo.

### 3.2 Protocolo experimental

El benchmark principal combina cinco datasets, cuatro modelos y diez semillas pareadas, dando lugar a 200 ejecuciones dataset-modelo-semilla. En los modelos profundos se fija un entrenamiento de 50 epochs, donde una epoch representa una pasada completa por los datos de entrenamiento; Gaussian Copula no usa epochs al no entrenarse como red neuronal. En cada combinación se separa entrenamiento/prueba, se entrena el generador sobre train, se genera un conjunto de datos sintéticos del mismo tamaño y se evalúa con el protocolo de la Tabla 1.

El uso de varias semillas permite comprobar si los resultados dependen de una ejecución concreta. La ablation de epochs modifica solo el número de epochs para observar cómo cambia el equilibrio entre aprendizaje, utilidad y privacidad empírica.

**Tabla 1:** Resumen del protocolo experimental.

| Elemento | Decisión experimental |
|---|---|
| Partición | Separación entrenamiento/prueba por dataset antes de entrenar generadores. |
| Tamaño del conjunto de datos sintéticos | Igual al tamaño del conjunto de entrenamiento real para evitar ventajas por volumen. |
| Semillas | Diez semillas pareadas en todas las combinaciones comparables. |
| Entrenamiento | 50 epochs en CTGAN, TVAE y TabDDPM; Gaussian Copula sin epochs. |
| Utilidad | TSTR: entrenar con datos sintéticos y evaluar en el conjunto de prueba real con RF y LR. |
| Fidelidad | SDMetrics, formas de columna, tendencias bivariadas y correlaciones. |
| Privacidad | Duplicados, NN train dist., sesgo train/test y categorías raras. |
| Ablation | 20, 50 y 100 epochs en CTGAN, TVAE y TabDDPM para los cinco datasets; Gaussian Copula queda fuera por no depender de epochs. |

La utilidad predictiva se mide con *Train on Synthetic, Test on Real* (TSTR): se entrena un clasificador con datos sintéticos y se evalúa en el test real. Los resultados se expresan como ratio frente al clasificador entrenado con datos reales; un valor cercano a 1 indica rendimiento similar y un valor superior a 1 indica una mejora puntual en esa métrica, no que el conjunto de datos sintéticos sea globalmente mejor. Se usan ROC-AUC, F1 y accuracy con Logistic Regression y Random Forest para cubrir clasificadores lineales y no lineales.

La fidelidad estadística se evalúa con el *Quality Report* de SDMetrics: *column shapes* mide similitud univariada y *column pair trends* mide relaciones bivariadas. La puntuación *overall quality* agrega ambas lecturas y resume el parecido distributivo entre el conjunto real y los datos sintéticos generados.

La privacidad se interpreta mediante un threat model acotado, inspirado en señales de pertenencia o proximidad al entrenamiento [10]. Se asume un atacante con acceso al conjunto de datos sintéticos y al esquema, capaz de buscar copias exactas o registros anormalmente próximos a registros reales. Se miden duplicados exactos, distancia media al vecino real más cercano del train y sesgo train/test. Este último no es una métrica separada del gap, sino una lectura conjunta: combina la fracción de registros de datos sintéticos más cercanos al train y el *train/test distance gap*, calculado como distancia media a test menos distancia media a train. También se mide la reproducción de categorías raras para detectar si el generador replica patrones minoritarios potencialmente identificables. Esta métrica solo se informa cuando el dataset contiene variables categóricas y suficientes categorías poco frecuentes; en escenarios donde no

aplica aparece como N/A. Las distancias NN se calculan en el espacio preprocesado del benchmark, por lo que se comparan dentro del mismo protocolo y no como distancias naturales universales.

El flujo experimental produjo salidas por semilla y agregaciones por dataset y modelo. Además, se calcularon contrastes pareados mediante la prueba de Wilcoxon pareada, comparando modelos sobre las mismas semillas y datasets. Los contrastes se usan como apoyo, ya que ayudan a identificar qué métricas muestran diferencias consistentes y cuáles fluctúan más entre semillas.

La reproducibilidad se apoya en dos decisiones: configuración explícita por dataset y modelo, y uso de semillas fijas. El tamaño del conjunto de datos sintéticos se igualó al tamaño del conjunto de entrenamiento real para evitar que una ventaja de utilidad o fidelidad viniera simplemente de generar más muestras.

## 4 RESULTADOS

La Tabla 2 resume cómo se lee cada dimensión antes de comparar modelos.

**Tabla 2:** Dimensiones de evaluación y lectura de métricas.

| Dimensión | Métrica principal | Pregunta que responde |
|---|---|---|
| Utilidad | Ratio RF AUC; F1 y accuracy como apoyo | ¿Sirven los datos sintéticos para entrenar modelos? |
| Fidelidad | Puntuación de calidad; shapes y pairs | ¿Se parece estadísticamente al real? |
| Privacidad | Duplicados, NN train dist. y sesgo train/test | ¿Hay copia o proximidad excesiva? |
| Estabilidad | Semillas y contrastes pareados | ¿El patrón depende de una ejecución? |

La Tabla 3 sintetiza la media global por modelo. La lectura principal es que los modelos no se ordenan igual según la dimensión evaluada. TVAE obtiene la mejor utilidad media con Random Forest, Gaussian Copula alcanza la mayor fidelidad global agregada, CTGAN mantiene duplicación prácticamente nula, y TabDDPM muestra la mayor distancia media al conjunto de entrenamiento, aunque con menor utilidad media y menor calidad global.

**Tabla 3:** Resumen global por modelo en las 200 ejecuciones.

| Modelo | RF ratio | Calidad | NN train dist. | Dup. |
|---|---|---|---|---|
| Gaussian Copula | 0,869 | **0,879** | 2,218 | $< 0{,}0001$ |
| CTGAN | 0,908 | 0,823 | 2,457 | $< 0{,}0001$ |
| TVAE | **0,965** | 0,868 | 0,805 | **0,0581** |
| TabDDPM | 0,801 | 0,740 | **8,151** | **0,0556** |

El análisis por dataset refuerza que la evaluación debe leerse por dimensión. En Adult y Bank Marketing, CTGAN preserva mejor la fidelidad global pero TVAE conserva mejor utilidad predictiva. En Default Credit, TabDDPM destaca en utilidad, mientras que Gaussian Copula preserva mejor la estructura estadística agregada. En Mushroom, TVAE obtiene la mayor utilidad pero no la mejor fidelidad global, y además presenta el problema de duplicación discutido más adelante.

### 4.1 Utilidad predictiva

En utilidad predictiva, TVAE obtiene los mejores resultados en la mayoría de datasets: lidera Adult (0,963), Bank Marketing (0,920), Dry Bean (0,992) y Mushroom (0,998) con Random Forest. TabDDPM lidera Default Credit (0,975), CTGAN resulta competitivo en Adult, Dry Bean y Mushroom, y Gaussian Copula conserva una utilidad razonable en Bank Marketing y Mushroom.

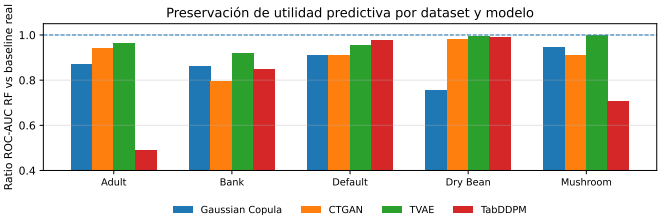


**Fig. 2:** Preservación de utilidad predictiva por dataset y modelo, medida como ratio de ROC-AUC frente al baseline real con Random Forest.

La Figura 2 muestra que la utilidad depende del dataset y no solo de la familia generativa. TabDDPM cae en Adult y Mushroom bajo el benchmark principal, pero es competitivo en Default Credit y Dry Bean; TVAE, en cambio, conserva señal predictiva de forma más estable.

El recuento agregado de contrastes pareados se resume en la Tabla A.2 del apéndice. Su lectura es de apoyo: indica en qué métricas las diferencias entre modelos se repiten con más frecuencia al comparar las mismas semillas y datasets. En este benchmark, las diferencias más frecuentes aparecen en utilidad con Random Forest y en distancia media al conjunto de entrenamiento.

### 4.2 Fidelidad estadística

En fidelidad, el patrón cambia. Gaussian Copula obtiene la mejor media global de *overall quality* (0,879), seguida de TVAE (0,868). Por dataset, CTGAN lidera Adult (0,846) y Bank Marketing (0,856), Gaussian Copula lidera Default Credit (0,905), TVAE lidera Dry Bean (0,917) y TabDDPM lidera Mushroom (0,935). El detalle por dataset y modelo se recoge en el Apéndice A.1, Figura A.1. La figura se centra únicamente en la puntuación de calidad, de forma que cada fila identifica explícitamente un modelo y la lectura visual no mezcla indicadores secundarios.

Esta diferencia entre utilidad y fidelidad es esperable. Un generador puede conservar bien distribuciones marginales y pares de columnas sin capturar exactamente la frontera predictiva útil para un clasificador. También puede ocurrir lo contrario: preservar la señal necesaria para el target, pero no todas las propiedades estadísticas del dataset original. El rendimiento competitivo de Gaussian Copula en calidad global sugiere que, para ciertas estructuras tabulares, un modelo estadístico bien ajustado puede capturar dependencias suficientes sin requerir una arquitectura profunda.

### 4.3 Privacidad empírica

La privacidad empírica se lee en tres niveles. Primero, la reproducción literal se observa mediante duplicados exactos. Segundo, la proximidad se resume con nearest-neighbour al conjunto de entrenamiento. Tercero, el sesgo train/test se usa como señal de apoyo: no es una métrica distinta del gap, sino la lectura conjunta de la fracción

de registros de datos sintéticos más cercanos al train y del *train/test distance gap*. Esta lectura permite detectar si los registros de datos sintéticos se aproximan sistemáticamente más al conjunto de entrenamiento que al conjunto de prueba. La matriz completa de privacidad se recoge en el Apéndice A.2, Figura A.2.

El caso Mushroom resulta especialmente ilustrativo. TVAE alcanza un RF ratio de 0,998, pero con una tasa media de duplicados exactos de 0,291. TabDDPM también presenta duplicación elevada en este dataset (0,278) con menor utilidad. En cambio, CTGAN y Gaussian Copula mantienen una tasa de duplicación reducida. Esto no implica que CTGAN sea siempre mejor, sino que Mushroom muestra un posible modo de fallo: en datasets categóricos muy estructurados, una utilidad elevada puede estar asociada a reproducción literal o casi literal de registros.

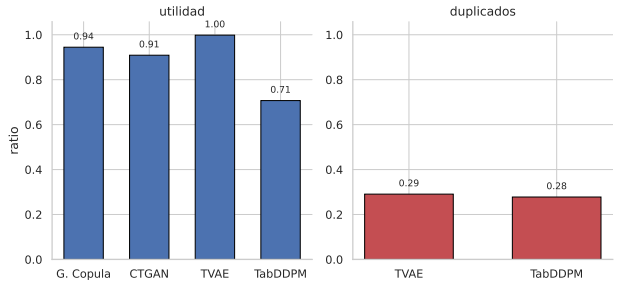


**Fig. 3:** caso Mushroom: comparación entre utilidad predictiva y proporción de duplicados exactos respecto al conjunto de entrenamiento. Ambos paneles usan la misma escala vertical; en duplicados se muestran solo los modelos con reproducción literal apreciable.

### 4.4 Ablation de epochs

El benchmark principal se mantiene en 50 epochs para comparar modelos bajo un presupuesto común. Además, se completó una ablation de epochs con 20, 50 y 100 epochs para CTGAN, TVAE y TabDDPM en los cinco datasets, usando tres semillas pareadas por combinación. Esto supone 135 ejecuciones de ablation. Gaussian Copula queda fuera de esta parte porque no se entrena mediante epochs.

La Tabla 4 resume los resultados agregados por modelo y número de epochs. En promedio, aumentar el entrenamiento mejora utilidad y fidelidad, pero también reduce la distancia media al conjunto de entrenamiento o aumenta la tasa de duplicados en algunos modelos. Por tanto, la ablation no sustituye el benchmark principal, sino que permite interpretar 50 epochs como un punto de referencia comparable y no como una configuración óptima universal.

**Tabla 4:** Ablation agregada por modelo y epochs. Promedios sobre cinco datasets y tres semillas; NN train dist. indica la distancia media al vecino real más cercano del conjunto de entrenamiento.

| Modelo | Epochs | RF ratio | Calidad | Dup. | NN train dist. |
|---|---|---|---|---|---|
| CTGAN | 20 | 0,726 | 0,775 | 0,000 | **2,83** |
| CTGAN | 50 | 0,900 | 0,821 | 0,000 | 2,46 |
| CTGAN | 100 | **0,942** | **0,861** | **0,003** | 1,95 |
| TVAE | 20 | 0,958 | 0,821 | **0,041** | **0,81** |
| TVAE | 50 | 0,964 | 0,866 | **0,057** | 0,81 |
| TVAE | 100 | **0,967** | **0,891** | **0,073** | 0,79 |
| TabDDPM | 20 | 0,671 | 0,665 | **0,039** | **14,16** |
| TabDDPM | 50 | 0,785 | 0,733 | **0,055** | 8,76 |
| TabDDPM | 100 | **0,960** | **0,855** | **0,089** | 3,79 |

La Figura 4 visualiza la evolución por epochs en el plano utilidad–fidelidad. El patrón más claro aparece en TabDDPM: a 100 epochs mejora mucho en utilidad y fidelidad, pero también se acerca al train y aumenta su duplicación media. TVAE ya parte de una utilidad alta en 20 epochs y mejora principalmente en fidelidad, mientras que CTGAN muestra una mejora progresiva con duplicación casi nula salvo una señal ligera a 100 epochs.

El detalle por dataset confirma que Mushroom es el caso donde la ablation resulta más crítica para privacidad: en TVAE y TabDDPM la utilidad se mantiene alta o aumenta, pero la duplicación exacta también crece con las epochs. La Tabla A.3 del apéndice recoge ese caso. Esto refuerza una conclusión transversal del estudio: maximizar utilidad o fidelidad sin observar privacidad empírica puede ocultar modos de fallo en datasets categóricos estructurados.

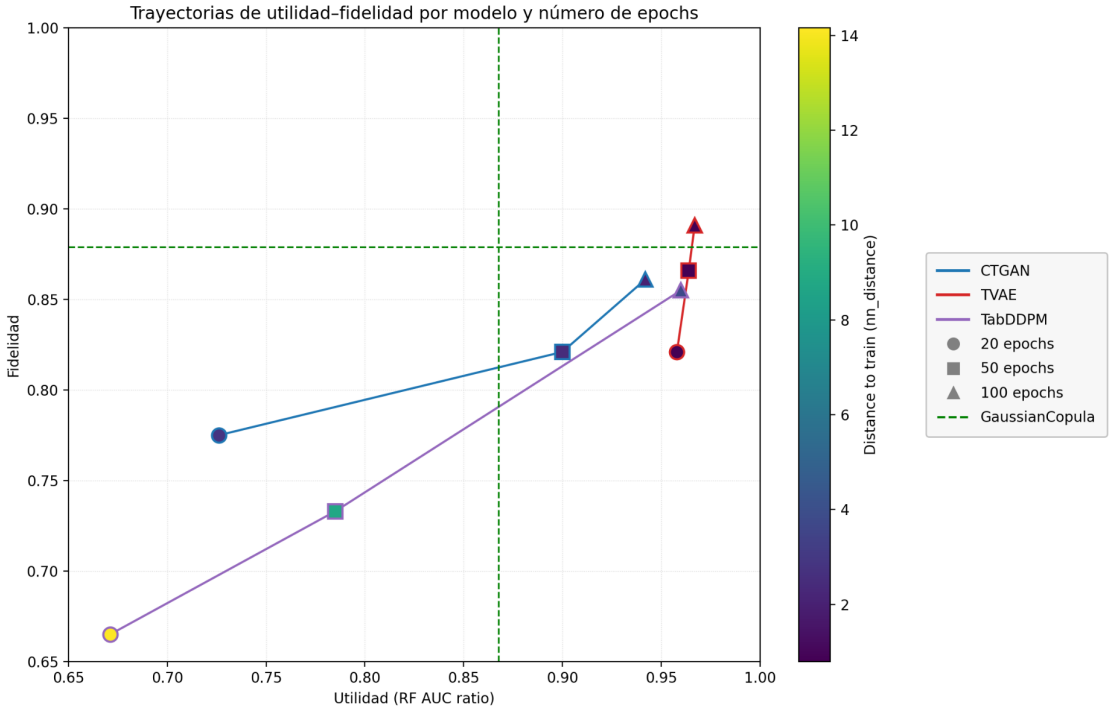


**Fig. 4:** Trayectorias de ablation entre 20, 50 y 100 epochs. El eje horizontal resume utilidad predictiva y el vertical fidelidad estadística; el color de la línea identifica el modelo, la forma del marcador identifica las epochs y la escala de color representa la distancia media al conjunto de entrenamiento. Las líneas verdes discontinuas muestran la referencia agregada de Gaussian Copula.

## 5 Discusión

La lectura más relevante no es solo qué modelo obtiene la mayor media, sino en qué tipo de dataset aparece cada comportamiento. Los retos tabulares condicionan la utilidad, la fidelidad y la privacidad empírica, y explican parte de las diferencias entre arquitecturas.

La Figura 5 resume que utilidad y proximidad al conjunto de entrenamiento no evolucionan siempre en la misma dirección, por lo que ambas dimensiones deben analizarse conjuntamente.

En Mushroom, dataset dominado por combinaciones categóricas muy cerradas, TVAE y TabDDPM alcanzan utilidad elevada o creciente, pero también reproducen más filas exactas. Este resultado sugiere que, cuando el espacio categórico contiene modos muy marcados, algunos generadores pueden obtener buen rendimiento copiando combinaciones frecuentes. En Default Credit, más numérico y con atípicos, TabDDPM resulta más competitivo, coherente con una arquitectura de difusión que puede beneficiarse de representar variación continua. Bank Marketing combina desbalance y variables de campaña, lo que afecta a la señal predictiva de CTGAN y TabDDPM bajo el presupuesto común. Dry Bean, al ser mayoritariamente numérico y más regular, permite ratios altos en varios modelos.

La comparación entre AUC y F1 en la Tabla A.4 refuerza esa lectura. En los escenarios más desbalanceados, algunos modelos conservan una AUC aceptable pero pierden más F1, señal de que el ordenamiento global de ejemplos se preserva mejor que el comportamiento sobre la clase minoritaria. Por ello, la utilidad no debe leerse solo como capacidad de ranking, sino también como capacidad para sostener decisiones útiles en escenarios con clases poco frecuentes.

La ablation confirma que 50 epochs funcionan como punto de comparación, no como techo de cada arquitectura. TabDDPM mejora de forma marcada a 100 epochs, pero esa mejora se acompaña de menor distancia al train y mayor duplicación media.

### 5.1 Síntesis de modelos

La Figura 6 cierra la lectura de resultados como guía cualitativa por modelo. No sustituye las tablas numéricas, sino que resume qué perfil resulta más conveniente según la dimensión priorizada.

La matriz debe leerse como una guía de decisión según la prioridad del caso de uso.

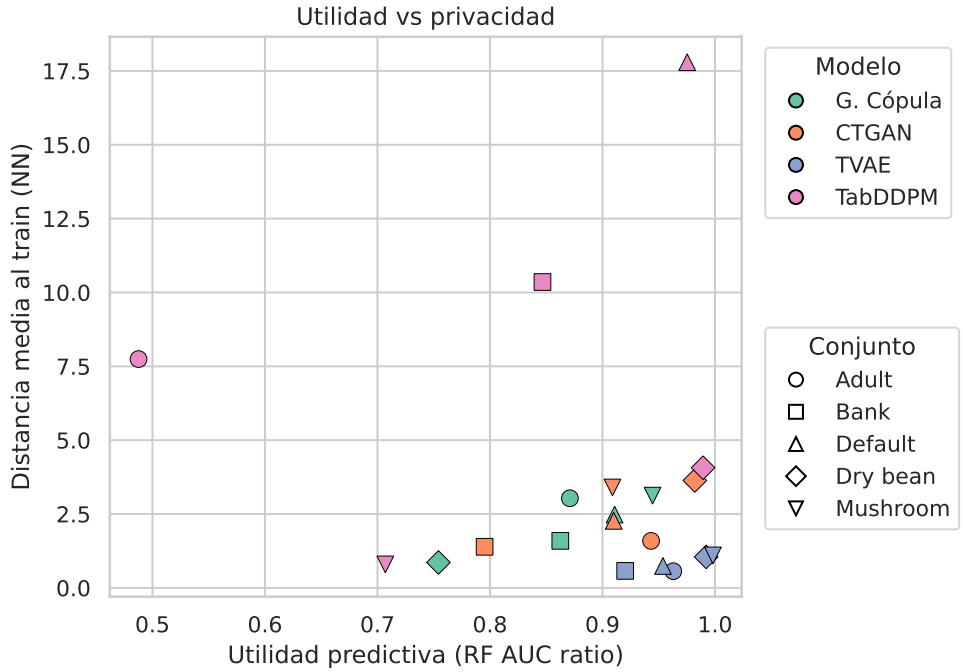


**Fig. 5:** Relación entre utilidad predictiva y distancia media al conjunto de entrenamiento para cada combinación dataset–modelo. El color identifica el modelo y la forma identifica el dataset; una mayor distancia no implica por sí sola privacidad formal.

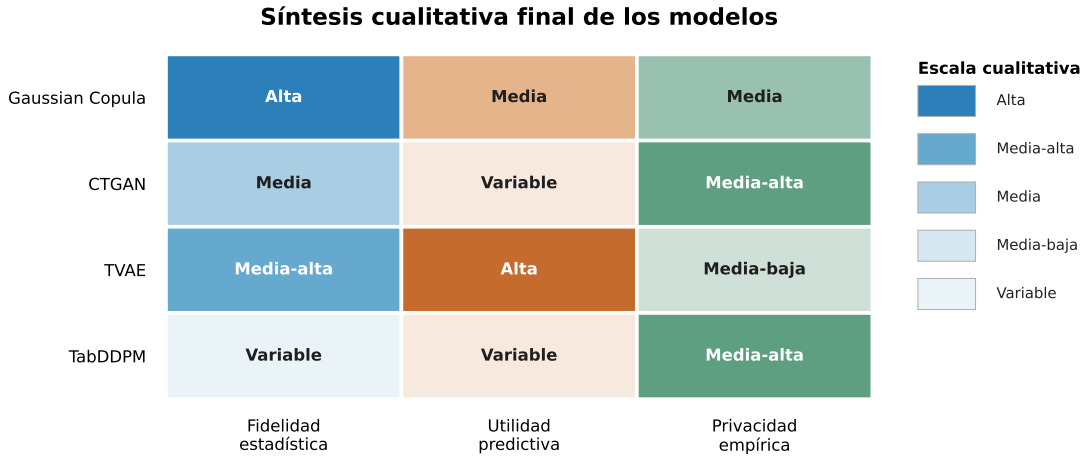


**Fig. 6:** Matriz final de lectura por modelo a partir de las métricas agregadas del benchmark.

## 6 Validez del estudio y limitaciones

La interpretación de los resultados debe considerar varias amenazas a la validez. La primera afecta a la validez interna: aunque todas las combinaciones usan el mismo protocolo, los modelos no tienen la misma naturaleza ni los mismos hiperparámetros. Gaussian Copula no entrena por epochs, mientras que CTGAN, TVAE y TabDDPM sí dependen de entrenamiento iterativo. Para mantener el benchmark comparable se fijaron configuraciones razonables, pero no se realizó una optimización exhaustiva por dataset. Por tanto, los resultados describen el comportamiento bajo un protocolo común, no el máximo rendimiento posible de cada modelo.

La segunda amenaza es la validez externa. Los cinco datasets cubren varios escenarios tabulares, pero siguen siendo datasets públicos relativamente controlados y de tamaño medio. En entornos reales pueden aparecer tablas con relaciones entre entidades, texto libre, variables temporales, identificadores, censura de datos, distribuciones cambiantes o reglas de negocio que no aparecen en el benchmark. Por ello, el estudio se limita a datos tabulares single-table.

La tercera amenaza afecta a las métricas. ROC-AUC, F1, accuracy, Quality Report y nearest-neighbour resumen dimensiones útiles, pero ninguna captura por sí sola toda la calidad de un conjunto de datos sintéticos. En utilidad, el resultado depende del clasificador usado; en fidelidad, los scores agregados pueden ocultar fallos locales; y en privacidad, las distancias dependen del preprocesado.

La cuarta amenaza se relaciona con la privacidad. Los indicadores usados detectan duplicados exactos, proximidad y categorías raras, pero no cubren adversarios más potentes. Además, un modelo alejado del entrenamiento podría parecer más privado simplemente porque genera datos de baja calidad, por lo que privacidad, utilidad y fidelidad

deben leerse juntas.

Por último, existe una limitación de recursos. El TFG prioriza un benchmark comparable y documentado frente a una búsqueda masiva de hiperparámetros. La ablation de epochs amplía el análisis, pero usa tres semillas por combinación, mientras que el benchmark principal usa diez; por ello debe interpretarse como análisis de sensibilidad y no como sustituto completo del protocolo principal. Esta decisión deja abierta una línea futura de optimización más profunda.

## 7 Conclusiones

El TFG ha construido y ejecutado un benchmark reproducible para evaluar datos sintéticos tabulares en cinco datasets públicos, cuatro familias de generadores y diez semillas. Con ello se cierran los objetivos principales: revisar modelos y herramientas, seleccionar datasets con retos distintos, definir un protocolo común, ejecutar los experimentos y analizar utilidad, fidelidad y privacidad empírica.

El hallazgo principal es que los modelos presentan perfiles complementarios. TVAE conserva mejor la utilidad media, Gaussian Copula y TVAE logran buena fidelidad agregada, CTGAN reduce la duplicación literal y TabDDPM puede mejorar con más entrenamiento, aunque con mayor variabilidad y mayor cercanía al train. Por tanto, la elección del generador depende del objetivo de uso y de la dimensión priorizada, no de una métrica única.

El análisis de privacidad confirma que los datos sintéticos no deben confundirse con anonimización formal. Duplicados exactos, nearest-neighbour y sesgo train/test detectan señales de memorización o proximidad excesiva, pero no sustituyen garantías formales ni ataques adversarios completos. En síntesis, el reto práctico no es solo generar datos parecidos, sino preservar señal útil y estructura estadística sin copiar en exceso el entrenamiento. Como trabajo futuro, se propone ampliar el benchmark a escenarios multi-tabla, incorporar membership inference más explícito, evaluar variantes con privacidad diferencial y profundizar en la sensibilidad de TabDDPM al preprocesado e hiperparámetros.

## Agradecimientos

Agradezco a Guillermo Navarro la tutorización del trabajo y las orientaciones para acotar el alcance del TFG. También agradezco el apoyo recibido durante el grado y las revisiones realizadas durante el desarrollo del informe.

## Referencias

[1] L. Xu, M. Skoularidou, A. Cuesta-Infante y K. Veeramachaneni, “Modeling tabular data using conditional GAN,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] D. P. Kingma y M. Welling, “Auto-Encoding Variational Bayes,” arXiv:1312.6114, 2013.

[3] A. Kotelnikov, D. Baranchuk, I. Rubachev y A. Babenko, “TabDDPM: Modelling Tabular Data with Diffusion Models,” *International Conference on Machine Learning*, 2023.

[4] J. Ho, A. Jain y P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[5] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen y A. Weller, “Synthetic data – what, why and how?” arXiv:2205.03257, 2022.

[6] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu y W. Wei, “Machine learning for synthetic data generation: a review,” arXiv:2302.04062, 2023.

[7] Synthetic Data Vault, “SDV documentation,” 2026. [En línea]. Disponible: `https://docs.sdv.dev/sdv`

[8] Synthetic Data Vault, “SDMetrics documentation,” 2026. [En línea]. Disponible: `https://docs.sdv.dev/sdmetrics`

[9] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. Springer, 2006.

[10] R. Shokri, M. Stronati, C. Song y V. Shmatikov, “Membership inference attacks against machine learning models,” *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

[11] B. Becker y R. Kohavi, “Adult,” UCI Machine Learning Repository. [En línea]. Disponible: `https://archive.ics.uci.edu/dataset/2/adult`

[12] S. Moro, P. Cortez y P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[13] I.-C. Yeh y C.-h. Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[14] M. Koklu y I. A. Ozkan, “Multiclass classification of dry beans using computer vision and machine learning techniques,” *Computers and Electronics in Agriculture*, vol. 174, 2020.

[15] J. Schlimmer, “Mushroom,” UCI Machine Learning Repository. [En línea]. Disponible: `https://archive.ics.uci.edu/dataset/73/mushroom`

[16] D. Dua y C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2017. [En línea]. Disponible: `https://archive.ics.uci.edu/`

# A Apéndice

## A.1 Fidelidad estadística por dataset y modelo

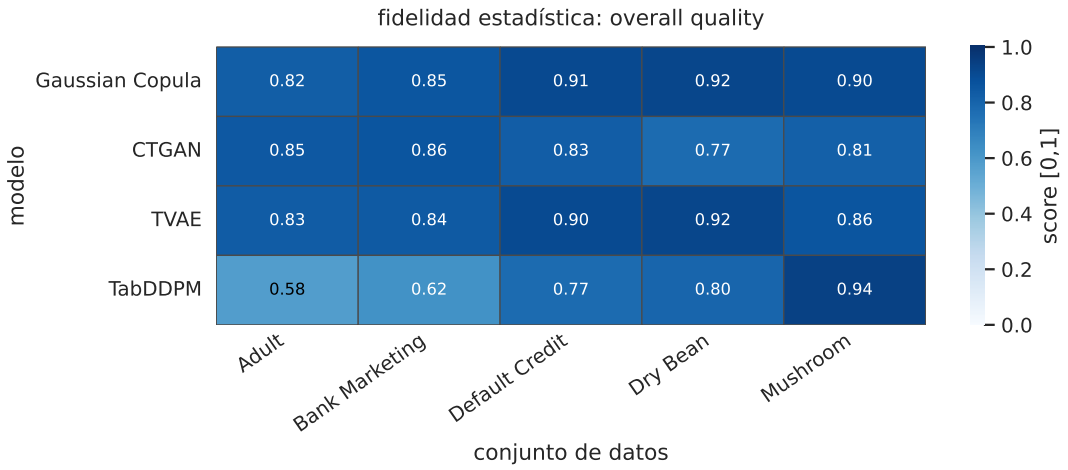


**Fig. A.1:** fidelidad estadística por modelo y dataset, medida con la puntuación *overall quality* de SDMetrics. Las filas identifican explícitamente cada modelo y las columnas corresponden a los cinco datasets del benchmark.

## A.2 Matriz completa de privacidad

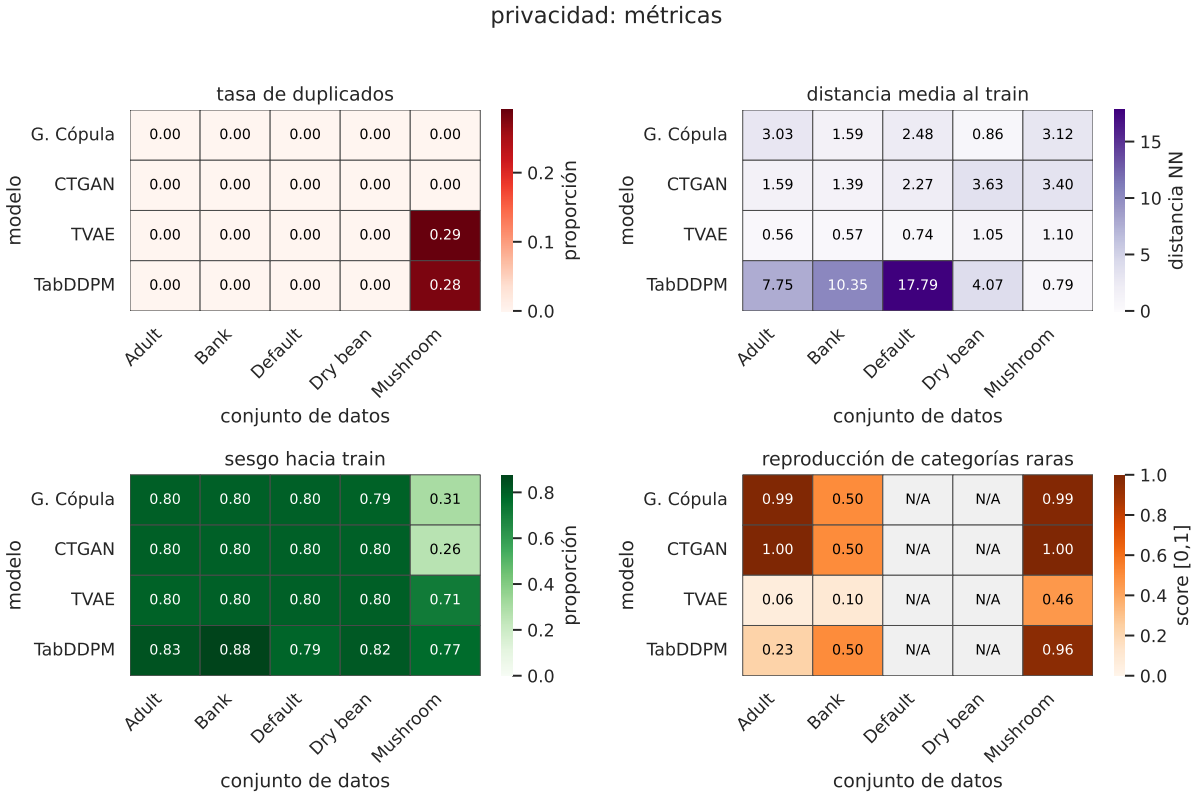


**Fig. A.2:** comparación de métricas de privacidad empírica por modelo y dataset: duplicados, distancia media al conjunto de entrenamiento, sesgo train/test y reproducción de categorías raras. Cuando esta última no aplica por ausencia de categorías raras relevantes, se muestra como N/A.

## A.3 Evolución del plan de trabajo

**Tabla A.1:** evolución del plan inicial durante el desarrollo del TFG.

| Rango aprox. | Plan inicial | Evolución real |
|---|---|---|
| 09/03–07/06 | Evaluar al menos dos modelos generativos. | Se compararon cuatro familias: Gaussian Copula, CTGAN, TVAE y TabDDPM. |
| 22/03–07/06 | Definir datasets durante el desarrollo. | Se consolidó un benchmark con Adult, Bank Marketing, Default Credit, Dry Bean y Mushroom. |
| 21/03–12/06 | Establecer métricas y protocolo de evaluación. | La evaluación se organizó en utilidad predictiva, fidelidad estadística y privacidad empírica. |
| 11/05–23/06 | Valorar ampliaciones si el tiempo lo permitía. | Se incorporaron perfilado de datasets, threat model, modelos adicionales y una ablation completa de epochs. |
| 08/06–23/06 | Ejecutar el estudio experimental principal y el análisis de sensibilidad. | Se completaron el benchmark principal y la ablation de epochs, manteniendo una trazabilidad separada de ejecuciones y resultados agregados. |
| 12/06–23/06 | Consolidar resultados para el dossier final. | Se agregaron métricas, se analizaron trade-offs y se prepararon visualizaciones de síntesis para el informe. |

## A.4 Soporte estadístico y sensibilidad en Mushroom

**Tabla A.2:** contrastes pareados significativos en el benchmark principal, calculados con prueba de Wilcoxon pareada. Cada métrica tiene 30 comparaciones posibles: cinco datasets por seis pares de modelos. La p representa la probabilidad de observar una diferencia al menos tan extrema si no hubiera diferencia sistemática entre los modelos comparados; se cuenta como significativa cuando $p < 0{,}05$.

| Métrica | Comparaciones | Significativas ($p < 0{,}05$) |
|---|---|---|
| NN train dist. | 30 | **28** |
| RF AUC ratio | 30 | 26 |
| LR AUC ratio | 30 | 25 |
| Calidad global | 30 | 25 |
| RF F1 ratio | 30 | 24 |
| Repr. categorías raras | 30 | 12 |
| Dup. rate | 30 | 5 |

**Tabla A.3:** detalle de Mushroom en la ablation de epochs. Se muestran los modelos donde la reproducción literal aumenta de forma más clara al incrementar el entrenamiento.

| Modelo | Epochs | RF ratio | Calidad | Dup. |
|---|---|---|---|---|
| TVAE | 20 | 0,992 | 0,763 | **0,206** |
| TVAE | 50 | 0,998 | 0,851 | **0,285** |
| TVAE | 100 | **1,000** | **0,916** | **0,365** |
| TabDDPM | 20 | 0,601 | 0,919 | **0,194** |
| TabDDPM | 50 | 0,704 | 0,931 | **0,276** |
| TabDDPM | 100 | **0,966** | **0,953** | **0,444** |

## A.5 Resumen numérico del benchmark principal

**Tabla A.4:** resumen numérico por combinación en el benchmark principal a 50 epochs. Los ratios comparan el rendimiento TSTR con el clasificador entrenado con datos reales; la abreviatura r. indica ratio.

| Dataset | Modelo | RF AUC r. | RF F1 r. | RF acc. r. | LR AUC r. | LR F1 r. | LR acc. r. | Calidad | NN train dist. | Dup. |
|---|---|---|---|---|---|---|---|---|---|---|
| Adult | GC | 0,871 | 0,081 | 0,898 | 0,935 | 0,145 | 0,904 | 0,823 | 3,03 | 0,000 |
| Adult | CTGAN | 0,943 | 0,873 | **0,965** | **0,957** | 0,911 | **0,966** | **0,846** | 1,59 | 0,000 |
| Adult | TVAE | **0,963** | **0,933** | 0,961 | 0,955 | **0,966** | 0,948 | 0,831 | 0,56 | $< 0{,}001$ |
| Adult | TabDDPM | 0,488 | 0,054 | 0,874 | 0,491 | 0,144 | 0,833 | 0,577 | **7,75** | 0,000 |
| Bank | GC | 0,863 | 0,002 | 0,976 | **0,935** | 0,115 | **0,981** | 0,855 | 1,59 | 0,000 |
| Bank | CTGAN | 0,795 | 0,604 | 0,959 | 0,808 | 0,684 | 0,970 | **0,856** | 1,39 | 0,000 |
| Bank | TVAE | **0,920** | **0,903** | 0,931 | 0,916 | **1,033** | 0,899 | 0,837 | 0,57 | 0,000 |
| Bank | TabDDPM | 0,847 | 0,123 | **0,977** | 0,735 | 0,027 | 0,981 | 0,623 | **10,35** | 0,000 |
| Default | GC | 0,911 | 0,261 | 0,956 | **1,001** | 0,144 | 0,967 | **0,905** | 2,48 | 0,000 |
| Default | CTGAN | 0,910 | **0,813** | 0,915 | 0,955 | **1,212** | 0,938 | 0,827 | 2,27 | 0,000 |
| Default | TVAE | 0,954 | 0,770 | **0,984** | 0,979 | 1,068 | **0,990** | 0,899 | 0,74 | 0,000 |
| Default | TabDDPM | **0,975** | 0,568 | 0,978 | 0,956 | 0,064 | 0,964 | 0,771 | **17,79** | 0,000 |
| DryBean | GC | 0,754 | 0,371 | 0,418 | 0,810 | 0,415 | 0,495 | 0,916 | 0,86 | 0,000 |
| DryBean | CTGAN | 0,982 | 0,826 | 0,823 | 0,983 | 0,830 | 0,845 | 0,772 | 3,63 | 0,000 |
| DryBean | TVAE | **0,992** | **0,952** | **0,961** | **0,996** | **0,955** | **0,964** | **0,917** | 1,05 | 0,000 |
| DryBean | TabDDPM | 0,989 | 0,812 | 0,874 | 0,993 | 0,790 | 0,832 | 0,796 | **4,07** | 0,000 |
| Mushroom | GC | 0,945 | 0,873 | 0,883 | 0,953 | 0,904 | 0,911 | 0,899 | 3,12 | $< 0{,}001$ |
| Mushroom | CTGAN | 0,909 | 0,753 | 0,799 | 0,918 | 0,841 | 0,858 | 0,814 | **3,40** | $< 0{,}001$ |
| Mushroom | TVAE | **0,998** | **0,974** | **0,975** | **0,996** | **0,976** | **0,977** | 0,856 | 1,10 | **0,291** |
| Mushroom | TabDDPM | 0,707 | 0,653 | 0,668 | 0,946 | 0,904 | 0,909 | **0,935** | 0,79 | **0,278** |