

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

<https://doi.org/10.35381/i.p.v8i15.5249>

Impacto de datos sintéticos generados mediante CTGAN en modelos de aprendizaje automático para la predicción

Impact of Synthetic Data Generated Using CTGAN on Machine Learning Models for Prediction

Wilmer Illescas-Espinoza

willescas@utmachala.edu.ec

Universidad Técnica de Machala, Machala, El Oro
Ecuador

<https://orcid.org/0000-0001-5171-7210>

Nancy Loja-Mora

nmloja@utmachala.edu.ec

Universidad Técnica de Machala, Machala, El Oro
Ecuador

<https://orcid.org/0000-0002-5583-4278>

Wilmer Rivas-Asanza

wrivas@utmachala.edu.ec

Universidad Técnica de Machala, Machala, El Oro
Ecuador

<https://orcid.org/0000-0002-2239-3664>

Recibido: 15 de junio 2026

Revisado: 10 de agosto 2026

Aprobado: 15 de septiembre 2026

Publicado: 01 de octubre 2026

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

RESUMEN

Este estudio tuvo como objetivo evaluar comparativamente el desempeño predictivo y la explicabilidad de Regresión Logística, Random Forest y XGBoost entrenados con datos reales frente a datos reales aumentados con CTGAN. La presente investigación es de naturaleza computacional/experimental sobre modelos, de nivel predictivo, con enfoque cuantitativo, propósito comparativo y orientación aplicada. Los resultados expresaron aspectos contrarios a lo que reporta la mayor parte de la literatura consultada, añadir datos sintéticos al entrenamiento no mejoró el desempeño predictivo de ninguno de los tres modelos, sino que lo redujo de forma estadísticamente significativa, incluso cuando esos mismos datos sintéticos superaron la validación estadística de fidelidad frente a los datos reales. En conclusión, a nivel teórico, el estudio aporta un caso concreto sobre la relación entre fidelidad estadística y utilidad predictiva de los datos sintéticos, un aspecto todavía poco explorado en la literatura revisada.

Descriptores: Inteligencia artificial; tratamiento de datos; análisis de datos; estadística; aprendizaje en línea. (Tesauro UNESCO)

ABSTRACT

The objective of this study was to evaluate comparatively the predictive performance and explainability of Logistic Regression, Random Forest, and XGBoost trained on real data versus real data augmented with CTGAN. This research is computational/experimental in nature, focusing on predictive models, with a quantitative approach, a comparative purpose, and an applied orientation. The results revealed findings contrary to what is reported in most of the literature reviewed: adding synthetic data to the training did not improve the predictive performance of any of the three models but rather reduced it in a statistically significant manner, even when that same synthetic data passed statistical fidelity validation compared to the real data. In conclusion, at the theoretical level, the study provides a concrete example of the relationship between statistical fidelity and the predictive utility of synthetic data an aspect that remains largely unexplored in the reviewed literature.

Descriptors: Artificial intelligence; data processing; data analysis; statistics; online learning. (UNESCO Thesaurus)

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

INTRODUCCIÓN

El sector financiero maneja información especialmente sensible, y por eso mismo conseguir datos reales, completos y bien balanceados para entrenar modelos de predicción de riesgo crediticio no es tan sencillo como parece: los historiales de incumplimiento suelen ser escasos, están fuertemente desbalanceados (la inmensa mayoría de los solicitantes sí paga) y muchas veces las instituciones no pueden compartirlos abiertamente por razones de privacidad y regulación. Esa restricción golpea directamente a los modelos de aprendizaje automático, que dependen de la cantidad y la calidad de los datos disponibles para generalizar bien.

Ante esa limitación, generar datos sintéticos (registros artificiales que imitan estadísticamente a los reales sin corresponder a ninguna persona concreta) se ha convertido en una alternativa cada vez más explorada para reforzar el entrenamiento sin comprometer la información de los solicitantes reales. Combinar generación sintética y aprendizaje automático en el ámbito financiero no es un tema nuevo, pero sí ha crecido de forma notoria en los últimos años. Al revisar la producción científica indexada en Scopus desde 2019 (año en que Xu et al. (2019) presentaron CTGAN) hasta la actualidad, el salto es evidente: de 31 artículos en 2023 a 94 en 2025, con 2026 acercándose ya a esa cifra sin que el año haya terminado.

Buena parte de esa producción se concentra en tres frentes: el manejo del desbalance de clases (SMOTE, sobremuestreo, submuestreo), la generación de datos sintéticos propiamente dicha (GAN, CTGAN, aumento de datos) y, de forma bastante más incipiente, la explicabilidad de los modelos resultantes (SHAP, LIME, XAI). Este último frente sigue siendo, comparado con los otros dos, notablemente más pequeño. Al revisar la literatura se evidencia que el corpus es joven y en franco ascenso investigativo, pues en 2023 hay 60 artículos científicos, 2025 llega a 94, y hasta el mes de septiembre del 2026 ya hay 78 documentos.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Cabe recalcar que de los 269 artículos revisados el 12.5% combinan con “credit” y de esos solo el 5.7% abordan la explicabilidad. Específicamente solo 7 artículos científicos tratan cuestiones de crédito con CTGAN y solo uno de los siete, el de Wang et al. (2026), considera la explicabilidad, aunque sin emplear SHAP (Adiputra et al., 2025; Li et al., 2025; Ramachandra y Vaithiyanathan, 2025; Zhao y Guan, 2023).

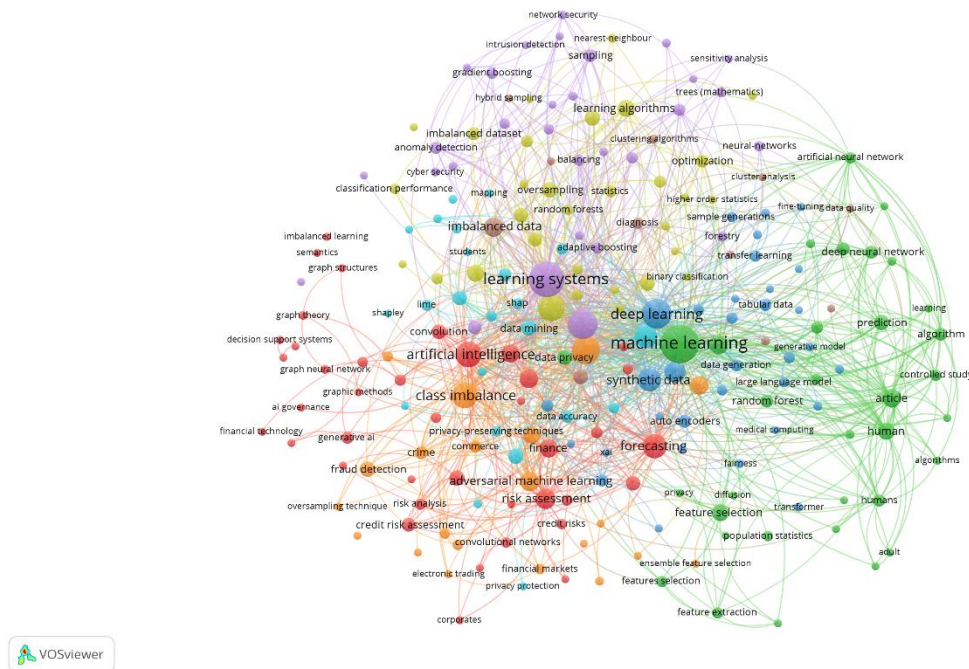


Figura 1. Mapa de coocurrencias de palabras claves.

Elaboración: Los autores.

Como se puede apreciar en la Figura 1, los artículos se dividen en seis comunidades temáticas como lo son de rojo el dominio específico tratado en este artículo, riesgo crediticio, finanzas, balance de clase; de naranja lo metodológico como datos sintéticos, generación de datos, forecasting y redes bayesianas; de púrpura técnicas y algoritmos en general; en amarillo manejo de desbalance de clases y estadística; verde Machine Learning genérico y cuestiones de la jerga médica; en turquesa, el núcleo, Machine

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Learning, Deep Learning, datos tabulares y random Forest. Se encuentra que el término recurrente fuera de las palabras claves, es el desbalance de clases, seguido por generative adversarial network debido a que CTGAN pertenece a aquella familia. Luego aparece la explicabilidad, pero como tema secundario, no central.

En cuanto a lo que aborda esta investigación (riesgo crediticio) es muy baja la presencia de investigaciones, por lo cual se evidencia que el estudio de la presente investigación aborda ese vacío científico. En cuanto al contexto científico de investigación se evidencia que los científicos más preocupados por estas temáticas son los asiáticos con 22 artículos, seguido por los europeos con 9, de los 39 que declararon su procedencia. De manera específica al analizar la afiliación se corrobora esta relación al encontrar 124 artículos de Asia (77 chinos, 30 hindúes, el resto de otros países de aquella región), y en nivel de producción preceden EEUU con 30 publicaciones y Reino Unido con 17 (entre los más relevantes).

Los experimentos o intervención son 56 artículos científicos, seguidos por 36 de regresión y econometría, y luego, en orden de relevancia, 15 de simulación y modelado. Entre los tres artículos más citados están Deep Neural Networks and Tabular Data: A Survey (779 citas) (Borisov et al., 2024), Domain Adaptation: Challenges, Methods, Datasets, and Applications (247 citas) (Singhal et al., 2023) y Deep transfer learning for automatic speech recognition: Towards better generalization (138 citas) (Kheddar et al., 2023).

Es importante precisar algunos términos. CTGAN pertenece a la familia GAN (red generativa antagónica), la cual consiste en dos redes neuronales que compiten entre sí durante el entrenamiento: una genera datos falsos y la otra intenta distinguirlos de los reales, hasta que la primera aprende a producir datos que la segunda ya no logra diferenciar.

Las iniciales CT de CTGAN provienen de que son generadores de datos sintéticos para datos que van en tablas (tabulares), y la C proviene del modo de muestreo, que es "condicional", es decir, generar datos sintéticos tiene sus particularidades, y no es lo

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

mismo generar datos para imágenes, por ejemplo, que para tablas con columnas numéricas y categóricas mezcladas. El riesgo crediticio, por su parte, es la probabilidad de que un solicitante no cumpla con sus obligaciones de pago; predecirlo con exactitud es, en el fondo, el problema que todos estos modelos intentan resolver. Y cuando se habla de explicabilidad, la referencia casi obligada es SHAP (Lundberg & Lee, 2017), una técnica que reparte, para cada predicción individual, cuánto "pesó" cada variable en el resultado final, algo que deja de ser opcional cuando un modelo decide si alguien recibe o no un crédito.

Esta investigación no cubre todo el universo de técnicas de generación sintética ni todos los algoritmos de clasificación disponibles. Se trabajó con un único conjunto de datos de riesgo crediticio (32,581 solicitudes), tres modelos: Regresión Logística, Random Forest y XGBoost, y una sola técnica de generación sintética, CTGAN, aplicada bajo una estrategia de aumento general: los datos sintéticos se incorporaron respetando la proporción de clases original del conjunto real, sin un sobremuestreo dirigido específicamente a los solicitantes que incumplen. La explicabilidad, a su vez, se evaluó únicamente con SHAP, dejando fuera otras técnicas como LIME. Estos límites no son un descuido, sino lo que permite mantener el estudio como una comparación controlada y reproducible entre dos escenarios de entrenamiento (solo datos reales, y datos reales más sintéticos).

No existe, en otras palabras, consenso científico sobre si los datos sintéticos generados con CTGAN mejoran, mantienen o perjudican la predicción de riesgo crediticio, y mucho menos sobre qué le ocurre a la explicabilidad del modelo cuando esos datos entran al entrenamiento. De ahí surge la pregunta que guía este estudio: ¿en qué medida la incorporación de datos sintéticos generados mediante CTGAN influye en el desempeño predictivo y la explicabilidad de los modelos de Regresión Logística, Random Forest y XGBoost para la predicción del riesgo crediticio, en comparación con el entrenamiento basado únicamente en datos reales?

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Con esa pregunta como eje, el objetivo de este estudio es evaluar comparativamente el desempeño predictivo y la explicabilidad de modelos de Regresión Logística, Random Forest y XGBoost entrenados con datos reales frente a los mismos modelos entrenados con datos reales aumentados mediante CTGAN, utilizando SHAP como técnica de explicación. La contribución de este trabajo no está en proponer una variante nueva de CTGAN ni un modelo de clasificación distinto, sino en algo más elemental que, según la revisión de literatura, nadie ha hecho todavía en este dominio específico: someter la combinación CTGAN + riesgo crediticio a una evaluación conjunta de desempeño y explicabilidad, respaldada por un protocolo estadístico (réplicas experimentales, pruebas bootstrap de significancia, validación de fidelidad estadística de los datos sintéticos) y no por una simple comparación puntual de métricas.

El resto del documento se organiza de la siguiente manera. En la metodología se detallan los métodos y técnicas utilizados para evaluar los algoritmos, en especial CTGAN; en los resultados se presentan las métricas de desempeño obtenidas en cada escenario junto con los hallazgos del análisis de explicabilidad con SHAP; en la discusión se contrastan esos resultados con lo reportado en la literatura revisada; y finalmente, en las conclusiones, se recogen los principales aportes, las limitaciones identificadas y las líneas de investigación que quedan abiertas.

MÉTODO

La presente investigación es de naturaleza computacional/experimental sobre modelos, de nivel predictivo, con enfoque cuantitativo, propósito comparativo y orientación aplicada. Se comparan dos escenarios de entrenamiento (datos reales y datos reales aumentados con datos sintéticos generados mediante CTGAN) para analizar su efecto sobre la predicción del riesgo crediticio, empleando Regresión Logística, Random Forest (Breiman, 2001) y XGBoost (Chen & Guestrin, 2016), y midiendo el desempeño a través de ROC-AUC, PR-AUC, Precisión, Recall y F1-Score. Por su temporalidad, el estudio es

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

transversal: los datos se analizan en un corte de tiempo determinado, sin seguimiento longitudinal.

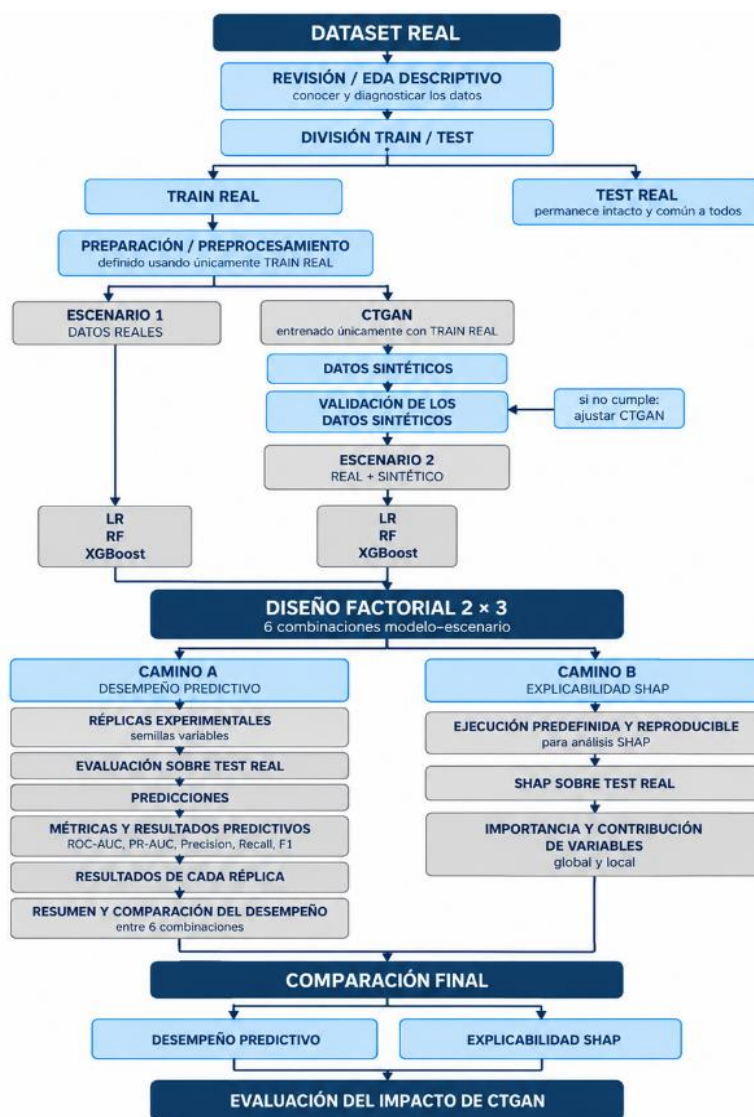


Figura 2. Diagrama de flujo de los procesos.
Elaboración: Los autores.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Se utilizó la base de datos abierta Credit Risk Dataset, disponible en el repositorio Kaggle (<https://www.kaggle.com/datasets/laotse/credit-risk-dataset>), descargada el 2026/09/09.

El estudio adopta un diseño factorial 2×3 : se cruzan dos escenarios de entrenamiento (datos reales / datos reales más sintéticos) con tres modelos de aprendizaje automático (Regresión Logística, Random Forest y XGBoost), lo que produce seis combinaciones modelo-escenario. Cada combinación se evalúa siguiendo dos caminos complementarios (véase Figura 2): el Camino A, orientado al desempeño predictivo, se ejecuta con 10 réplicas variando la semilla del modelo; el Camino B, orientado a la explicabilidad, se ejecuta con una semilla fija (42) sobre una única corrida reproducible.

Se plantean las siguientes hipótesis, evaluadas con un nivel de significancia de 0.05:

H0: la incorporación de datos sintéticos al conjunto de entrenamiento no produce diferencias significativas en el desempeño predictivo de los modelos respecto al escenario con datos reales.

H1: la incorporación de datos sintéticos sí produce diferencias significativas en el desempeño predictivo de los modelos respecto al escenario con datos reales.

Para la revisión de literatura se empleó la estrategia PICOC, formulada originalmente por Petticrew y Roberts (2006) como una adaptación del PICO propio de la medicina basada en evidencia, incorporando el elemento de Contexto; se complementó con las guías de Petersen et al. (2008) para el desarrollo de mapping studies en ingeniería de software. Se definieron los siguientes componentes: Población (P) "credit risk", "credit scoring", "credit default"; Intervención (I) "synthetic data", "synthetic data generation", CTGAN, "conditional tabular GAN"; Contexto (C) "machine learning", "classification". Se decidió omitir los elementos de Comparación y Resultados (Outcome), por considerarse que su inclusión podía sobre-refinar la búsqueda y excluir artículos relevantes en un campo todavía emergente. Se consideraron artículos publicados a partir de 2022, en idioma inglés, en las bases Scopus y Web of Science (WOS). La cadena de búsqueda resultante fue:

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

("synthetic data" OR "synthetic data generation" OR CTGAN OR "conditional tabular GAN")

AND ("credit risk" OR "credit scoring" OR "credit default")

AND ("machine learning" OR "classification")

AND (LIMIT-TO(DOCTYPE, "ar")) AND (LIMIT-TO(LANGUAGE, "English"))

AND PUBYEAR > 2022 AND PUBYEAR < 2027

Esta búsqueda arrojó 263 artículos en Scopus y 6 en WOS, los cuales fueron depurados eliminando duplicados, literatura gris y documentos en español, antes de su análisis. Para el procesamiento del corpus exportado desde Scopus (formato .csv) se utilizó el software VOSviewer (van Eck & Waltman, 2010).

El procesamiento de los datos siguió una secuencia fija: primero un análisis exploratorio (EDA) para diagnosticar la calidad y estructura del conjunto; después, la división en conjuntos de entrenamiento (TRAIN) y prueba (TEST); y finalmente, la definición del preprocesamiento (imputación, codificación, escalado) ajustada únicamente sobre TRAIN, aplicando después las mismas reglas (sin reajustarlas) sobre TEST, que permanece reservado para la evaluación final. Solo una vez completados estos tres pasos se establecieron las condiciones experimentales del estudio.

Para la construcción del Escenario 2, los datos sintéticos generados mediante CTGAN se sumaron íntegramente al conjunto de entrenamiento real, sin eliminar ninguna clase. CTGAN fue entrenado sobre ambas clases (pagadores y no pagadores) del conjunto real, por lo que el lote sintético generado replica, de forma aproximada, la misma proporción de clases observada en los datos reales. Es decir, la estrategia de aumento de datos no estuvo dirigida específicamente a la clase minoritaria (en la programación, ESTRATEGIA_CLASE = "general").

Cada réplica del Camino A permite observar la estabilidad de las predicciones ante variaciones en la semilla del modelo; el análisis SHAP, en cambio, expone la estructura

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

interna que explica cada predicción y el peso relativo de cada variable, por lo que se realiza sobre una única ejecución fija y reproducible, en lugar de repetirse por réplica.

Para la validación estadística de los datos sintéticos, los tres primeros intentos se evaluaron con la prueba de Kolmogorov-Smirnov (KS) bajo un criterio de significancia de $p\text{-valor} = 0.05$, sin que los datos sintéticos lograran pasar la validación. Al examinar este resultado, se identificó que el tamaño de la muestra (~26,000 observaciones) era la causa: la literatura documenta que, con muestras grandes, el $p\text{-valor}$ se vuelve una función excesivamente sensible al tamaño muestral, generando el rechazo sistemático de H_0 incluso ante diferencias sin relevancia práctica entre las distribuciones comparadas (Gómez-de-Mariscal et al., 2021). Por esta razón, se decidió adoptar como criterio de validación el estadístico KS de forma directa (la distancia máxima entre las funciones de distribución acumuladas), fijando un umbral de aceptación práctico de 0.17.

RESULTADOS

Caracterización general de los datos

El conjunto de datos original contenía 32,581 solicitudes de crédito. Durante la revisión exploratoria se detectaron valores faltantes en `person_emp_length` (895 casos) y `loan_int_rate` (3,116 casos), imputados posteriormente con la mediana calculada sobre TRAIN. También se identificaron valores físicamente imposibles (edades de hasta 144 años y antigüedades laborales de 123 años) que fueron eliminados del conjunto, junto con 165 registros duplicados (96 de ellos en TRAIN). Tras la partición en TRAIN (80%) y TEST (20%), el conjunto de entrenamiento quedó conformado por 25,961 observaciones, con una proporción de clases de 78.2% (pagan) frente a 21.8% (incumplen).

Validación estadística de los datos sintéticos

CTGAN, entrenado durante 300 épocas exclusivamente sobre TRAIN, valor predeterminado de la implementación oficial de CTGAN (Xu et al., 2019) generó un lote

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

sintético del mismo tamaño (25,961 filas). Las siete variables numéricas evaluadas mediante la prueba de Kolmogorov-Smirnov se ubicaron por debajo del umbral de aceptación (0.17), confirmando una similitud estadística aceptable entre las distribuciones real y sintética. El caso más cercano al límite fue `loan_int_rate`, con un estadístico de 0.168. Al combinar TRAIN con el lote sintético completo (estrategia general), el Escenario 2 quedó conformado por 51,922 filas, con una proporción de clases de 70.9%/29.1%.

Camino A: desempeño predictivo

La Tabla 1 resume el desempeño promedio (10 réplicas) de los tres modelos en ambos escenarios, evaluados sobre el mismo conjunto de prueba.

Tabla 1.

Resumen de las réplicas con los tres modelos ML.

Modelo	Escenario	ROC-AUC	PR-AUC	Precisión	Recall	F1-Score
Regresión Logística	1 (real)	0.866	0.7141	0.749	0.5162	0.6112
Regresión Logística	2 (real+sintético)	0.8557	0.7	0.6919	0.5781	0.6299
Random Forest	1 (real)	0.9327	0.8848	0.98	0.7169	0.8281
Random Forest	2 (real+sintético)	0.9235	0.8712	0.9034	0.7243	0.804
XGBoost	1 (real)	0.9493	0.9061	0.9584	0.7361	0.8327
XGBoost	2 (real+sintético)	0.9335	0.8827	0.9106	0.7273	0.8087

Elaboración: Los autores.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

En los tres modelos, el ROC-AUC y el PR-AUC disminuyeron al pasar del Escenario 1 al Escenario 2. La comparación bootstrap (2,000 iteraciones, IC 95%) confirmó que esta disminución fue estadísticamente significativa en los tres casos, según se detalla en la Tabla 2.

Tabla 2.
Comparación Bootstrap.

Modelo	Diferencia ROC-AUC (E2-E1)	IC 95%	¿Diferencia significativa?
Regresión Logística	-0.0103	[-0.0103, -0.0103]	Sí
Random Forest	-0.0091	[-0.0096, -0.0086]	Sí
XGBoost	-0.0157	[-0.0166, -0.0149]	Sí

Elaboración: Los autores.

Vale precisar un comportamiento distinto en el caso de Regresión Logística: mientras que en Random Forest y XGBoost el F1-Score también disminuyó entre escenarios, en Regresión Logística ocurrió lo contrario (el F1 subió de 0.6112 a 0.6299), sostenido por un aumento del recall (0.5162 a 0.5781) mayor a la caída de la precisión (0.7490 a 0.6919).

Camino B: explicabilidad (SHAP)

La estabilidad del ranking de importancia de variables entre escenarios, medida mediante la correlación de Spearman, resultó distinta según el tipo de modelo (Tabla 3).

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Tabla 3.
Correlación Spearman.

Modelo	Correlación de Spearman (E1 vs. E2)
Regresión Logística	0.6421
Random Forest	0.8917
XGBoost	0.9323

Elaboración: Los autores.

Se puede evidenciar que los modelos basados en árboles (RF y XGBoost) mantuvieron rankings de importancia SHAP muy similares entre escenarios, mientras que RL al ser un modelo paramétrico mostró una reorganización más marcada, como se evidencia en la Tabla 3, donde los modelos de árboles presentan coeficientes más cercanos a 1. Esto se debe a que los modelos tipo árbol generan distribución de datos de forma no paramétrica, lo cual los hace menos sensibles a cambios en la distribución de los datos sintéticos, en cambio, los modelos lineales, dependen directamente de esa distribución. Las Tablas 4 y 5 presentan las cinco variables con mayor importancia SHAP promedio (valor absoluto medio) por modelo, en cada escenario.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Tabla 4.
Escenario 1 (datos reales).

Orden	Regresión Logística	Random Forest	XGBoost
1	loan_percent_income (1.046)	loan_percent_income (0.082)	person_income (1.132)
2	loan_grade (0.953)	loan_grade (0.081)	loan_grade (1.029)
3	loan_amnt (0.500)	person_income (0.051)	loan_percent_income (0.787)
4	person_home_ownership_RENT (0.293)	person_home_ownership_RENT (0.043)	person_home_ownership_OW N (0.450)
5	person_home_ownership_OW N (0.260)	loan_int_rate (0.030)	person_home_ownership_RE NT (0.364)

Elaboración: Los autores.

En la Tabla 4 observamos como en RL las variables dominantes son loan_percent_income y loan_grade, esto tiene sentido pues al ser un modelo lineal se apoya fuerte en pocas variables con relación directa y fuerte con el objetivo. En RF observamos que son las mismas variables las dominantes, pero más bajos en magnitud, porque SHAP reparte la contribución entre más variables, además aparece loan_int_rate. En XGBoost es el que más difiere poniendo la variable person_income como la más dominante y aparecen dos variables dummy OWN y RENT.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Tabla 5.
 Escenario 2 (real + sintético).

Orden	Regresión Logística	Random Forest	XGBoost
1	loan_percent_income (0.649)	loan_percent_income (0.079)	person_income (0.764)
2	loan_grade (0.553)	loan_grade (0.079)	loan_percent_income (0.613)
3	person_income (0.320)	person_income (0.063)	loan_grade (0.606)
4	person_home_ownership_RENT (0.314)	person_home_ownership_RENT (0.056)	loan_int_rate (0.429)
5	loan_int_rate (0.313)	loan_int_rate (0.051)	person_home_ownership_RENT (0.375)

Elaboración: Los autores.

En ambos escenarios y en los tres modelos, loan_percent_income, loan_grade y person_income se ubican de forma consistente entre las variables más relevantes, aunque el orden exacto varía según el modelo y el escenario.

DISCUSIÓN

Sobre la caída del desempeño predictivo al incorporar datos sintéticos

El hallazgo central de este estudio (que la incorporación de datos sintéticos generados con CTGAN redujo, de forma estadísticamente significativa, el ROC-AUC y el PR-AUC de los tres modelos evaluados) se aparta de la tendencia dominante en la literatura revisada. De los ocho estudios de la matriz que evalúan directamente el efecto de la generación sintética sobre el desempeño predictivo en contextos crediticios, ocho reportan mejoras (Raheem et al., 2024; Pourkhoshgoftar et al., 2025; Xueping et al., 2026; Shan et al., 2026; Li et al., 2025), frente a un único estudio que documenta un resultado compatible con el nuestro.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Esta diferencia numérica (8 a 1) no debería entenderse, sin más, como evidencia de que nuestro resultado esté equivocado. Una diferencia metodológica relevante separa a este estudio de la mayoría de los antecedentes citados: la generación sintética aquí se aplicó bajo una estrategia general (CTGAN entrenado sobre ambas clases, replicando aproximadamente la proporción real (78 % / 22 %)) y no bajo un sobremuestreo dirigido específicamente a la clase minoritaria, que es la práctica habitual en los estudios que reportan mejoras. Bajo la estrategia empleada aquí, duplicar el conjunto de entrenamiento no corrige el desbalance de forma tan directa como sí lo hacen SMOTE, ADASYN o CTGAN aplicado exclusivamente a la clase de incumplimiento. Esta diferencia de diseño ofrece una explicación plausible (aunque no comprobada dentro del alcance de este estudio) de por qué el efecto observado difiere del reportado en gran parte de la literatura, y constituye una de las líneas de verificación pendientes señaladas en las limitaciones.

Sobre la jerarquía de desempeño entre modelos

El orden de desempeño observado (XGBoost por encima de Random Forest, y ambos por encima de la Regresión Logística, en los dos escenarios) es consistente con lo poco que la literatura revisada permite comparar en este punto específico. El Sayad et al. (2026) documentan una superioridad similar de XGBoost frente a otros modelos de aprendizaje automático en la evaluación de riesgo hipotecario, y Adiputra et al. (2025) reportan que las mejores combinaciones surgen al emparejar métodos de generación sintética con modelos de ensamble antes que con clasificadores clásicos. Se trata, sin embargo, de una comparación con evidencia todavía escasa en el corpus revisado (solo dos estudios comparables, ninguno con resultados contrarios), por lo que este hallazgo se sostiene más en la coherencia con el comportamiento típico de estos algoritmos frente a datos tabulares desbalanceados que en un respaldo bibliográfico extenso.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Sobre la relación entre tipo de modelo y estabilidad de la explicabilidad

La estabilidad del ranking de importancia SHAP entre escenarios fue notablemente distinta según la familia de modelo: alta en Random Forest ($p = 0.8917$) y XGBoost ($p = 0.9323$), y solo moderada en Regresión Logística ($p = 0.6421$). No fue posible contrastar este hallazgo específico contra la literatura, dado que ninguno de los 263 documentos revisados evalúa la estabilidad de la explicabilidad de un mismo modelo entre un escenario de entrenamiento real y otro aumentado con datos sintéticos. Este vacío es consistente con lo ya documentado en el análisis bibliométrico del estado del arte: de los siete estudios que combinan CTGAN con riesgo crediticio, ninguno incorpora una técnica de explicabilidad de modelo. La diferencia observada entre modelos lineales y modelos de árboles podría estar asociada a la forma en que cada familia de algoritmos traduce cambios en la distribución de los datos de entrenamiento en cambios de sus parámetros internos (los coeficientes de una regresión logística responden de forma más directa a desplazamientos en la distribución conjunta de las variables que las particiones de un árbol), pero esta es una interpretación razonable a partir del resultado observado, no una relación causal verificada por este estudio.

Sobre la fidelidad estadística frente a la utilidad predictiva

Uno de los hallazgos con menor correspondencia en la literatura revisada es que los datos sintéticos superaron la validación de fidelidad estadística (Kolmogorov-Smirnov, umbral de 0.17, en las siete variables numéricas evaluadas) al mismo tiempo que el desempeño predictivo de los tres modelos disminuyó de forma significativa. Ningún estudio del corpus reporta conjuntamente una prueba de bondad de ajuste y su efecto sobre el desempeño de un modelo predictivo en riesgo crediticio, lo que impide una comparación directa. El hallazgo sugiere, dentro de los límites de este estudio, que una prueba de similitud distribucional univariada (columna por columna) no necesariamente captura si las relaciones conjuntas entre variables que un modelo predictivo aprende del

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

entrenamiento se conservan en los datos sintéticos. Esta distinción entre fidelidad estadística y utilidad predictiva no ha sido, hasta donde permite establecer la revisión realizada, examinada de forma explícita en el dominio de riesgo crediticio.

Sobre las variables más relevantes para el riesgo crediticio

Las variables `loan_percent_income`, `loan_grade` y `person_income` se mantuvieron consistentemente entre las de mayor peso según SHAP, en los tres modelos y en ambos escenarios. Los dos estudios de la matriz que sí reportan una jerarquía explícita de variables importantes identifican predictores distintos: Gao y Zhou (2026) señalan razones financieras corporativas (retorno sobre activos, cuentas por cobrar), y Ali Shahee y Patel (2025) señalan el estado de la cuenta corriente como principal predictor, con la tasa de interés como coincidencia parcial. Esta divergencia es, en gran medida, esperable: ambos estudios operan sobre un objeto de predicción distinto (riesgo corporativo e historiales crediticios de tipo europeo (German Credit Data), respectivamente) por lo que no constituye evidencia contraria a nuestro resultado, sino una comparación entre poblaciones y unidades de análisis diferentes.

Sobre el comportamiento distinto de la Regresión Logística

A diferencia de Random Forest y XGBoost, donde Precision, Recall y F1 disminuyeron de forma conjunta en el Escenario 2, la Regresión Logística mostró un patrón mixto: la Precisión cayó (0.7490 a 0.6919), pero el Recall aumentó (0.5162 a 0.5781) lo suficiente como para que el F1 subiera (0.6112 a 0.6299). El único estudio comparable de la matriz que reporta explícitamente el comportamiento conjunto de estas tres métricas documenta que su método de aumento de datos mejoró las cuatro métricas simultáneamente, sin el desplazamiento diferenciado observado aquí. La ausencia de más estudios que desglosen este comportamiento por tipo de modelo limita la posibilidad de una comparación más sólida en este punto.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

CONCLUSIONES

Cumplimiento de los objetivos. El objetivo general (evaluar comparativamente el desempeño predictivo y la explicabilidad de Regresión Logística, Random Forest y XGBoost entrenados con datos reales frente a datos reales aumentados con CTGAN, usando SHAP como técnica de explicación) se cumplió en su totalidad. De los cuatro objetivos específicos planteados, el estado del arte y la selección del conjunto de datos, la evaluación del Escenario 1, la generación y evaluación del Escenario 2 y la comparación final entre ambos escenarios se ejecutaron conforme a lo previsto en el diseño factorial 2×3 .

Síntesis de los hallazgos. Contrario a lo que reporta la mayor parte de la literatura consultada, añadir datos sintéticos al entrenamiento no mejoró el desempeño predictivo de ninguno de los tres modelos, sino que lo redujo de forma estadísticamente significativa, incluso cuando esos mismos datos sintéticos superaron la validación estadística de fidelidad frente a los datos reales. La explicabilidad SHAP se mantuvo prácticamente estable en los modelos de árboles ante ese mismo cambio, pero varió de forma más marcada en el modelo lineal. Un hallazgo adicional, no anticipado, fue que la Regresión Logística mostró un comportamiento distinto al de los otros dos modelos en el balance entre precisión y recall al incorporar los datos sintéticos.

Implicaciones. Estos resultados indican que la conveniencia de aumentar datos de entrenamiento con CTGAN no puede asumirse de forma automática en el dominio de riesgo crediticio, y que superar una prueba de fidelidad estadística no garantiza una mejora (ni siquiera la ausencia de un deterioro) en la tarea predictiva final. A nivel metodológico, el estudio muestra también que la estabilidad de la explicabilidad ante el aumento de datos no es uniforme entre familias de modelos, lo que sugiere que evaluar únicamente el desempeño, sin acompañarlo de un análisis de explicabilidad, puede ocultar efectos relevantes para la adopción de estos modelos en contextos regulados.

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Relación con la investigación existente. Los hallazgos de este estudio contrastan con la mayoría de los antecedentes que evalúan el efecto de la generación sintética sobre el desempeño predictivo en crédito, y no encuentran contraparte directa en la literatura en lo referente a la estabilidad de la explicabilidad SHAP ante el aumento de datos ni a la relación entre fidelidad estadística y utilidad predictiva, ambos aspectos coherentes con el vacío de conocimiento identificado en la revisión bibliométrica, donde ninguno de los estudios que combinan CTGAN y riesgo crediticio incorpora una técnica de explicabilidad de modelo. El desarrollo completo de estas comparaciones se aborda en la sección de Discusión.

Contribuciones. Este trabajo aporta, en el plano metodológico, un protocolo experimental replicable para evaluar el efecto de datos sintéticos que integra réplicas experimentales, pruebas de significancia mediante bootstrap y validación estadística de fidelidad de los datos generados. En el plano empírico, aporta evidencia (en un dominio y bajo una estrategia de aumento específicos) de que la generación sintética con CTGAN puede perjudicar antes que beneficiar el desempeño predictivo, y de que ese efecto no es uniforme entre modelos ni entre las dimensiones de desempeño y explicabilidad. En el plano práctico, los resultados llaman a la cautela frente a la adopción de datos sintéticos en el desarrollo de modelos de scoring crediticio sin una validación específica de su efecto sobre la tarea predictiva final.

Limitaciones y líneas futuras. El estudio se limita a un único conjunto de datos y a una única técnica de generación sintética, aplicada además bajo una estrategia de aumento general y no dirigida a la clase minoritaria; la validación de fidelidad de los datos sintéticos, por su parte, se realizó únicamente sobre las variables numéricas, sin extenderse a las variables categóricas del conjunto de datos. Para trabajos futuros se recomienda comparar más técnicas de generación de datos sintéticos tabulares, como VAE, GAN, SMOTE, CopulaGAN o redes bayesianas, así como aplicar CTGAN bajo una estrategia dirigida específicamente a la clase minoritaria para contrastar los hallazgos

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

aquí obtenidos. Se recomienda también estudiar qué proporción de datos sintéticos resulta más conveniente para mejorar los modelos predictivos, y extender la validación de fidelidad estadística a las variables categóricas del conjunto de datos.

Aplicaciones. A nivel práctico, los hallazgos son relevantes para instituciones financieras que evalúan incorporar datos sintéticos en el desarrollo de modelos de scoring crediticio. A nivel metodológico, el diseño factorial 2×3 con evaluación dual de desempeño y explicabilidad puede aplicarse a otras técnicas de generación sintética o a otros dominios de clasificación con datos desbalanceados. A nivel teórico, el estudio aporta un caso concreto sobre la relación entre fidelidad estadística y utilidad predictiva de los datos sintéticos, un aspecto todavía poco explorado en la literatura revisada.

CONFLICTO DE INTERÉS

Los autores declaran que no tienen conflicto de interés en la publicación de este artículo.

FINANCIAMIENTO

No monetario.

AGRADECIMIENTO

A todos los actores involucrados por su aporte y valioso apoyo en el desarrollo de esta investigación.

REFERENCIAS CONSULTADAS

- Adiputra, I. N. M., Lin, P.-C., & Wanchai, P. (2025). The effectiveness of generative adversarial network-based oversampling methods for imbalanced multi-class credit score classification. *Electronics*, 14(4), Article 697. <https://doi.org/10.3390/electronics14040697>
- Ali Shahee, S., & Patel, R. (2025). An explainable ADASYN-based focal loss approach for credit assessment. *Journal of Forecasting*, 44(4), 1513–1530. <https://doi.org/10.1002/for.3252>

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

- Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7499–7519. <https://n9.cl/khip9>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- El Sayad, I., Chong, F. H. L., & Gosal, B. (2026). Risk assessment in mortgages: A comparative study of AI models. *PeerJ Computer Science*, 12, Article e3494. <https://doi.org/10.7717/peerj-cs.3494>
- Gao, X., & Zhou, Y. (2026). A resampling ensemble model for multi-window corporate default prediction under class imbalance. *Systems*, 14(7), Article 776. <https://n9.cl/nk6n55>
- Gómez-de-Mariscal, E., Guerrero, V., Sneider, A., Jayatilaka, H., Phillip, J. M., Wirtz, D., & Muñoz-Barrutia, A. (2021). Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. *Scientific Reports*, 11, Article 20942. <https://n9.cl/je6yl>
- Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems*, 277, Article 110851. <https://doi.org/10.1016/j.knosys.2023.110851>
- Li, Z., Chen, G., Liu, Y., & Yuan, J. (2025). An efficient multi-institution collaborative learning framework based on knowledge filtering and data sharing. *Cluster Computing*, 28(10), Article 644. <https://n9.cl/wmjvdd>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://n9.cl/82snlp>

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

- Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. En 12th International Conference on Evaluation and Assessment in Software Engineering (EASE). BCS Learning & Development. <https://doi.org/10.14236/ewic/EASE2008.8>
- Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Blackwell Publishing. <https://n9.cl/rw7s1r>
- Pourkhoshgoftar, S., Shahbahrami, A., & Esmi, N. (2025). Graph-based inductive learning for credit risk prediction with imbalance mitigation. Computational Economics. <https://doi.org/10.1007/s10614-025-11114-9>
- Raheem, E. A., Dinar, A. M., Mohammed, M. A., & Al-Attar, B. (2024). Improving loan status prediction accuracy with generative adversarial networks: Addressing data scarcity and bias. Journal of Intelligent Systems and Internet of Things, 13(1), 251–258. <https://doi.org/10.54216/JISIoT.130118>
- Ramachandra, P., & Vaithiyanathan, S. (2025). Fed-DPSDG-WGAN: Differentially private synthetic data generation for loan default prediction via federated Wasserstein GAN. IEEE Access, 13, 52069–52084. <https://n9.cl/knhnf>
- Shan, W., Kang, X., & Gao, B. (2026). A TabPFN-based framework for credit risk prediction in automotive green supply chain finance. Sustainability, 18(12), Article 6305. <https://doi.org/10.3390/su18126305>
- Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Domain adaptation: Challenges, methods, datasets, and applications. IEEE Access, 11, 6973–7020. <https://n9.cl/3wvi76>
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32. <https://n9.cl/rfe5on>
- Xueping, Z., Samuri, S. M., & Adnan, M. H. M. (2026). Adaptive anomaly correction and dynamic sample balancing for deep intelligent credit risk assessment framework. NED University Journal of Research, 23(1), 44–69. <https://doi.org/10.35453/NEDJR-ASCN-2025-007.R4>

Wilmer Illescas-Espinoza; Nancy Loja-Mora; Wilmer Rivas-Asanza

Zhao, X., & Guan, S. (2023). CTCN: A novel credit card fraud detection method based on Conditional Tabular Generative Adversarial Networks and Temporal Convolutional Network. PeerJ Computer Science, 9, Article e1634. <https://doi.org/10.7717/peerj-cs.1634>

©2026 por los autores. Este artículo es de acceso abierto y distribuido según los términos y condiciones de la licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0) (<https://creativecommons.org/licenses/by-nc-sa/4.0/>).