Sobre Métodos de IA Generativa (SOTA 2024-2025):

TVAE vs. CTGAN: "En datasets tabulares de menos de 1,000 registros, ¿por qué se considera que el Tabular Variational Autoencoder (TVAE) es más estable y preserva mejor la diversidad que las Redes Generativas Antagónicas (CTGAN)?"

ForestFlow: "¿Cómo funciona el algoritmo ForestFlow (basado en XGBoost) para generar datos sintéticos y cuáles son sus ventajas en tareas de regresión con variables técnicas escasas?"

TabSyn (Diffusion Models): "¿Es viable aplicar modelos de Difusión Latente (como TabSyn) para aumentar datos de encuestas universitarias preservando las correlaciones no lineales entre habilidades blandas y técnicas?"

Sobre Métodos Estadísticos y de Regresión: SMOGN para Regresión: "¿Cómo mejora el algoritmo SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) el rendimiento de modelos XGBoost cuando el tiempo de inserción laboral está sesgado hacia el valor cero?"

Data Augmentation via LLMs: "¿Cuál es la efectividad de usar LLMs (Large Language Models) como 'TabuLa' para generar perfiles sintéticos de graduados a partir de vectores de competencias técnicas?"

Sobre la Evaluación de Calidad: Fidelidad vs. Utilidad: "¿Qué métricas científicas (como el Hellinger Distance o el ML Utility Score) son necesarias para demostrar a un tribunal que los datos aumentados no introdujeron sesgos artificiales en la tesis?"

Métodos SOTA de datos sintéticos para encuestas pequeñas en 2024-2025

1. TVAE vs. CTGAN en n<1000

En tabulares pequeños, **TVAE suele ser más estable y diverso que CTGAN** porque:

- **Entrenamiento**: VAE optimiza una ELBO con objetivo bien definido; esto reduce el colapso de modos típico de GANs y la sensibilidad extrema a hiperparámetros que se observa en CTGAN (Miletic & Sariyar, 2024).
- **Robustez y estabilidad**: evaluaciones comparativas en microdatos tabulares muestran que CTGAN es muy sensible a épocas y batch size, mientras que TVAE mantiene utilidad alta con menos ajuste fino (Miletic & Sariyar, 2024).
- **Cobertura de la distribución**: al muestrear del espacio latente regularizado, TVAE tiende a cubrir mejor modos raros; GANs priorizan “realismo local” y pueden ignorar regiones poco frecuentes, reduciendo diversidad en n pequeño (Miletic & Sariyar, 2024; Zhang et al., 2023).

En síntesis, en muestras <1000, la combinación de objetivo probabilístico y menor sensibilidad a hiperparámetros hace que TVAE sea más predecible y que preserve mejor la variabilidad.

Comparativa CTGAN–TVAE (propiedades globales)

Modelo	Estabilidad (datasets)	Sensibilidad a hiperparámetros	Utilidad media (ML)	Comentario	Citas
CTGAN	Variable	Alta	Media	Buen manejo de categóricas, pero inestable	(Miletic & Sariyar, 2024; Zhang et al., 2023; Alqulaity & Yang, 2024)
TVAE	Alta	Baja-media	Alta	Robusto incluso en alta dimensión y n limitado	(Miletic & Sariyar, 2024; Zhang et al., 2023)

FIGURE 1 Comparación cualitativa CTGAN vs TVAE en estabilidad y utilidad.

2. ForestFlow (basado en XGBoost)

Funcionamiento (visión rápida)
ForestFlow (FF) es un modelo de **flow matching condicional para datos tabulares**:

1. Define una familia de distribuciones intermedias que llevan una gaussiana estándar a la distribución real.
2. En lugar de redes neuronales, usa **regresores XGBoost** para aprender el **campo de velocidades condicional** (derivada de los datos respecto al tiempo del flujo) (Akazan et al., 2024).
3. Integra una ODE (p.ej. Euler) desde ruido gaussiano hasta datos sintéticos, usando las predicciones de XGBoost en cada paso (Akazan et al., 2024).

Ventajas en regresión con pocas variables técnicas

- **No requiere GPU**; funciona muy bien con 12–24 cores CPU, útil en entornos académicos con recursos limitados (Akazan et al., 2024).
- XGBoost **aprende bien relaciones no lineales e interacciones** aun con dimensionalidad moderada y pocos predictores “duros”, generando tablas con fuerte **utilidad predictiva aguas abajo** (Akazan et al., 2024).
- Comparaciones amplias muestran que ForestFlow **supera o iguala a GAN/VAE profundos** en calidad y diversidad de datos tabulares en múltiples datasets (Akazan et al., 2024).

Limitaciones: trata categóricas con one-hot continuo, lo que introduce errores y sensibilidad a condiciones iniciales; HS3F propone una versión secuencial más rápida y robusta (Akazan et al., 2024).

3. Modelos de difusión latente (TabSyn y afines)

TabSyn aplica un **VAE para mapear la tabla a un espacio latente continuo** y entrena allí un **modelo de difusión score-based** que:

- Unifica numéricas y categóricas en un único espacio latente.
- Optimiza explícitamente la **distribución de embeddings** para facilitar la difusión.
- Luego decodifica de vuelta a tabla (Zhang et al., 2023).

Resultados en seis datasets muestran que:

- Reduce errores en **distribuciones por columna** y **correlaciones entre columnas** hasta un 86% y 67% frente a los mejores baselines, incluyendo CTGAN y TVAE (Zhang et al., 2023).
- Supera a otros métodos (GAN, VAE, difusión directa, LLMs tipo GReaT) en métricas de correlación y calidad global (Zhang et al., 2023).

Conclusión práctica: sí, **modelos de difusión latente como TabSyn son viables para ampliar encuestas universitarias** preservando **correlaciones no lineales entre habilidades blandas y técnicas**, siempre que se disponga de suficientes registros para entrenar el VAE latente (habitualmente > unos cientos de filas) (Zhang et al., 2023; Zhang, 2025).

4. SMOGN + XGBoost en regresión sesgada a cero

SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) crea **muestras sintéticas en regiones poco representadas del espacio continuo** (p.ej., tiempos de inserción laboral largos) añadiendo ruido gaussiano controlado.

En un estudio de regresión con solo 110 observaciones, combinar **SMOGN + XGBoost**, más clustering y tuning de hiperparámetros, mejoró notablemente el rendimiento:

- MSE ≈ 0.0002 y $R^2 \approx 0.98$ en test, frente a configuraciones sin SMOGN considerablemente peores (Krop et al., 2024).

Mecanismo relevante para tu caso (tiempos muy concentrados en 0):

- Re-pondera la pérdida hacia las colas (tiempos largos).
- Genera alrededor de estos puntos sintéticos con ruido suave, para que XGBoost no aprenda solo el pico en 0 (Krop et al., 2024).

Aplicado a "tiempo de inserción → sesgo a 0", SMOGN ayuda a **aprender mejor la cola derecha** (egresados que tardan mucho), mejorando métricas como RMSE y MAE en esos casos extremos.

5. LLMs tipo TabuLa para datos tabulares sintéticos

La literatura sobre LLMs tabulares está emergiendo. Dos líneas relevantes:

- **GReaT / TabuLa-like**: formatean la tabla como texto y usan LLM autoregresivo para generar filas; estudios recientes encuentran que LLMs pueden competir con modelos clásicos de tablas en generación sintética, pero la calidad depende fuertemente del prompting y del tamaño del modelo (Zhang et al., 2023; Barr et al., 2025).
- Un estudio 2025 muestra que **GPT-4o puede generar datos tabulares sintéticos "zero-shot"**, preservando medias, intervalos de confianza y especialmente **correlaciones bivariadas (dirección y magnitud)** mejor que CTGAN en tres datasets pequeños, aunque con limitaciones en la forma precisa de las distribuciones (Barr et al., 2025).

Para "perfiles sintéticos de graduados a partir de vectores de competencias técnicas":

- Los LLMs parecen **eficaces para preservar correlaciones y resúmenes estadísticos básicos** incluso sin fine-tuning (Barr et al., 2025).
- Sin embargo, pueden distorsionar colas y distribuciones no normales; se recomiendan transformaciones (log, estandarización) y pasar al modelo estadísticas detalladas (rango, skewness) para controlar esto (Barr et al., 2025).

En comparación con TVAE/TabSyn, los LLMs ofrecen **gran flexibilidad semántica**, pero **menos garantías probabilísticas**; conviene usarlos como complemento más que sustituto en una tesis metodológicamente defensible.

6. Métricas de fidelidad vs. utilidad para convencer a un tribunal

La evaluación de datos sintéticos se basa en dos ejes:

6.1 Fidelidad estadística

Para demostrar que no se han introducido sesgos artificiales:

- **Distancias de distribución**:
  - **Hellinger** o **Jensen-Shannon** entre distribuciones marginales y conjuntas (reales vs sintéticas) (Zhang et al., 2023; Miletic & Sariyar, 2024).
- **Tests por columna**: KS para numéricas; χ^2 o Jaccard para categóricas (Alqulaity & Yang, 2024; Miletic & Sariyar, 2024).
- **Correlaciones**:
  - Diferencia absoluta en matrices de correlación (Pearson/Spearman, o mutua información para no lineales) real vs sintética; TabSyn usa errores de correlación como métrica clave (Zhang et al., 2023; Zhang, 2025).

6.2 Utilidad para modelos de ML

- **ML Utility Score / Train-on-Synthetic, Test-on-Real (TSTR)**: entrenar XGBoost (u otro) en sintético y evaluar en real; si la performance (R^2, RMSE, AUC) se mantiene cercana, se evidencia alta utilidad sin distorsión grave de patrones (Zhang et al., 2023; Ara et al., 2025; Miletic & Sariyar, 2024).
- **Train-on-Real, Test-on-Synthetic (TRTS)**: evalúa si los datos sintéticos cubren bien el soporte de los reales (Miletic & Sariyar, 2024).

6.3 Equidad y ausencia de nuevos sesgos

- Analizar **rendimiento de modelos (o distribuciones)** por subgrupos (género, área de estudio, estrato socioeconómico). Estudios recientes muestran que incluso modelos muy precisos pueden tener fuertes sesgos por subgrupo; la corrección se puede hacer regularizando la correlación entre errores y atributos protegidos (Ugirumurera et al., 2024).
- Mostrar que las métricas de equidad (p.ej., diferencias de error o predicción media por grupo) **no empeoran** al usar datos aumentados (Ugirumurera et al., 2024).

Tabla resumen de métricas clave

Objetivo	Métricas recomendadas	Uso en defensa de tesis	Citas
Fidelidad marginal	Hellinger / JS, KS, χ^2, Jaccard	"Las distribuciones originales no se han deformado"	(Zhang et al., 2023; Alqulaity & Yang, 2024; Miletic & Sariyar, 2024)
Fidelidad relacional	Error en matrices de correlación / MI	"Se preservan relaciones blandas–técnicas, lineales y no lineales"	(Zhang et al., 2023; Zhang, 2025; Barr et al., 2025)
Utilidad ML	TSTR/TRTS con XGBoost, métricas de predicción	"Los modelos basados en sintético generalizan al mundo real"	(Zhang et al., 2023; Ara et al., 2025; Miletic & Sariyar, 2024)
Equidad	Diferencias de error por grupo; correlación error–atributo protegido	"No se introducen nuevos sesgos grupales"	(Ugirumurera et al., 2024; Miletic & Sariyar, 2024)

FIGURE 2 Métricas clave para evaluar fidelidad, utilidad y equidad de datos sintéticos.

7. Cómo articularlo en una tesis de inserción laboral universitaria

1. **Metodología de generación**
   - Explicar por qué se eligen **TVAE / TabSyn / ForestFlow** frente a CTGAN (estabilidad, robustez en n pequeño, preservación de correlaciones) (Zhang et al., 2023; Miletic & Sariyar, 2024; Akazan et al., 2024).
   - Justificar SMOGN para corregir la cola de tiempos de inserción laboral (Krop et al., 2024).
   - Si se usan LLMs (tipo TabuLa), delimitar claramente su rol (p.ej., generación de perfiles de ejemplo o escenarios de simulación), y no como único generador de la base (Barr et al., 2025).
2. **Protocolo de evaluación**
   - Diseñar una batería estándar: distancias marginales (Hellinger/KS), fidelidad de correlaciones (Pearson/Spearman/MI), TSTR con XGBoost, y análisis de sesgos por subgrupo (Zhang et al., 2023; Ara et al., 2025; Ugirumurera et al., 2024; Miletic & Sariyar, 2024; Barr et al., 2025).
   - Documentar umbrales razonables (p.ej., diferencias de correlación <0.05–0.10, degradación de R^2 <5–10%).
3. **Transparencia y reproducibilidad**
   - Reportar hiperparámetros y semillas, enfatizando que métodos como TVAE y ForestFlow muestran menor sensibilidad que CTGAN en distintos datasets (Miletic & Sariyar, 2024; Akazan et al., 2024).
   - Incluir apéndice con figuras de distribución real vs sintética y resultados TSTR/TRTS.

Con este esquema, se puede argumentar ante un tribunal que el aumento de datos **mejora la capacidad explicativa y predictiva** de los modelos de inserción laboral **sin introducir sesgos artificiales y respetando la estructura estadística y relacional de la muestra original**.

These papers were sourced and synthesized using Consensus, an AI-powered search engine for research. Try it at https://consensus.app

References

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., & Karypis, G. (2023). Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. *ArXiv*, abs/2310.09656. https://doi.org/10.48550/arxiv.2310.09656

Ara, S., R, T., & S.H., M. (2025). Predictive Model to Analyse Real and Synthetic Data for Learners' Performance Prediction Using Regression Techniques. *Online Learning*. https://doi.org/10.24059/olj.v29i1.4390

Krop, I., Sasaoka, T., Shimada, H., & Hamanaka, A. (2024). Optimizing Mean Fragment Size Prediction in Rock Blasting: A Synergistic Approach Combining Clustering, Hyperparameter Tuning, and Data Augmentation. *Eng*. https://doi.org/10.3390/eng5030102

Alqulaity, M., & Yang, P. (2024). Enhanced Conditional GAN for High-Quality Synthetic Tabular Data Generation in Mobile-Based Cardiovascular Healthcare. *Sensors (Basel, Switzerland)*, 24. https://doi.org/10.3390/s24237673

Zhang, Q. (2025). Latent Diffusion Models with Correlation Loss for Tabular Data Generation. *2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 1849-1853. https://doi.org/10.1109/nnice64954.2025.11064643

Ugirumurera, J., Bensen, E., Severino, J., & Sanyal, J. (2024). Addressing bias in bagging and boosting regression models. *Scientific Reports*, 14. https://doi.org/10.1038/s41598-024-68907-5

Miletic, M., & Sariyar, M. (2024). Challenges of Using Synthetic Data Generation Methods for Tabular Microdata. *Applied Sciences*. https://doi.org/10.3390/app14145975

Barr, A., Rozman, R., & Guo, E. (2025). Generative adversarial networks vs large language models: a comparative study on synthetic tabular data generation. *ArXiv*, abs/2502.14523. https://doi.org/10.48550/arxiv.2502.14523

Akazan, A., Jolicoeur-Martineau, A., & Mitliagkas, I. (2024). Generating Tabular Data Using Heterogeneous Sequential Feature Forest Flow Matching. *ArXiv*, abs/2410.15516. https://doi.org/10.48550/arxiv.2410.15516