Datos Sintéticos:
Introducción a Técnicas Generativas y Evaluación de Calidad

Diego Andrés Cleves Leguízamo

Facultad de Estadística, Universidad Santo Tomás-Bogotá

Opción de Grado

Javier Mauricio Sierra

22 de Noviembre de 2025

Resumen

El presente trabajo aborda el estudio de datos sintéticos desde su concepción teórica y generación. Se propone la implementación de diversos modelos con el fin de sintetizar datos categóricos y cuantitativos, luego se comparan de acuerdo a su capacidad de enmascarar datos (propensión), su medidas de semenjanza estadística y tiempo de ejecución. Los resutados mostraron que simular variables categóricas con base a reglas, que representan sus dependencias en la realidad es el mejor método para simularlas. No obstante, a las variables numéricas no fue posible sintetizarlas de manera adecuada, los modelos propuestos no capturaron la cópula adecuadamente. A manera de conclusión se indica dónde se fallo y las oportunidades de mejora disponibles.

Palabras clave: sintetización, semejanza, propensión, dependencia, cópula

Abstract

This work studies synthetic data from its theoretical conception to its generation. Several models are implemented to synthesize categorical and numerical data and are compared in terms of data masking capability (propensity), statistical similarity, and execution time. The results indicate that rule-based simulation is the most effective approach for categorical variables, while numerical variables could not be adequately synthesized due to the models' inability to capture the copula structure. The conclusions discuss the identified limitations and potential improvements.

Keywords: synthesis, similarity, propensity, dependence, copula

Introducción

La proposición y desarrollo de metodologías para generar datos sintéticos no son nuevas, de hecho su proposición inicial data de hace más de 40 años. El catalizador para ello fue la creciente demanda por el respeto a la intimidad tanto por parte del Estado como de las empresas privadas, que podría resumirse en la acción jurisdiccional conocida como *Habeas Data*(el sujeto es poseedor de sus datos, en español). Ni la Estadística ni la Ciencias de la Computación fueron ajenas a estos fenómenos sociales, aunque se optó por usar métodos de enmascaramiento más sencillos de ejecutar que la replicación de bases de datos.

Desde la Estadística, en 1978 Ronald Rubin insinuó que sus técnicas de imputación múltiple podrían ser usadas para enmascarar datos, aunque no refinó su propuesta teórica hasta 1993. Liew y otros científicos computacionales, en 1985, propusieron métodos para remplazar datos sensibles por datos sintéticos. Su motivación fue la creciente exigencia por parte de las autoridades estadounidenses de proteger la privacidad de los consumidores, que se manifestó en las distintas leyes de protección de datos que rigen a cada Estado, aunque el precedente judicial más antiguo del siglo XX se encuentra en el artículo 12 de la Declaración de los Derechos Humanos, mas su motivación no fue comercial sino más bien política [1].

En la actualidad la motivación por desarrollar metodologías para generarlos ha sido el desarrollo de los LLMs (Large languages models, en inglés), SMLs(small languages models, en inglés) y de tratamientos médicos, ya que necesitan gran cantidad de datos de alta calidad para el entrenamiento de sus modelos, los cuales suelen ser escasos. Los datos sintéticos se presentan como una solución a la carencia de datos reales útiles. Por ejemplo: Phi-4 y DeepSeek V3 fueron entrenados con estos, y NVidia desarrolló Nemotron, una familia de modelos que generan datos sintenticos para entrenar LLMs.

Para proseguir al desarrollo del tema, se indica que la exposición del caso de estudio que realiza este trabajo se divide en las siguientes secciones: (I) Marco teórico, (II) análisis

[1] Si se desea conocer más sobre el tema, se le sugiere consultar sobre las Leyes de Nuremberg, los experimentos humanos realizados por el *tercer Reich* y el Imperio Japonés.

descriptivo y exploratorio de los datos reales, (III) simulación de variables categóricas, (IV) simulación de variables numéricas, (V) Resultados y (VI) conclusiones.

1. Marco teórico

1.1. Estado del arte

Ninguna idea de la que parte está exposición es original, por eso, en aras de reconocer el trabajo de quienes han dedicado sus vidas a la investigación del tema que nos convoca, en esta sección se presentarán los resumenes de algunas investigaciones que resultaron ser útiles para el desarrollo de este trabajo. Se espera que también les sean útiles, apreciado lector.

1.1.1. Estudios fundacionales

Los primeros desarrollos teóricos y prácticos de la generación de datos sintéticos se encontraron en los estudios de Ronald Rubin y Liew, publicados en 1993 y 1985, respectivamente. El primero se titula *Discussion: Statistical Disclosure Limitation*, en él Rubin teoriza sobre las posibilidades, sin ignorar las limitaciones, de este tipo de datos. En el segundo, titulado *A Data Distortion by Probability Distribution*, Liew y otros, proponen métodos de generación, evaluación de los datos generados y de su propensión a ser diferenciados de los reales. Muchas de las inquietudes de Rubin sobre esta propuesta se siguen discutiendo y los métodos propuestos por Liew aún se usan.

Un trabajo incluso más antiguo, que no fue considerado importante en los inicios de esta disciplina, pero habría de serlo, es *A Mathematical Theory of Communication* de Claude Shannon, que publicó en 1948, propuso una medida de la información dentro de un sistema, lo cual es útil al momento de comparar los datos reales y los sintéticos en función de la información que poseen: se espera que los datos tengan una información similar. También fue la base para la divergencia de Jensen y Shannon, medida de semejanza estadística usada durante esta investigación.

1.1.2. Estudios sobre medidas de propensión y utilidad

Estos estudios han procurado resolver el problema del compromiso entre la privacidad y la utilidad. Tiancheng Li y Ninghui Li, en su investigación titulada *On the tradeoff between privacy and utility in data publishing*, publicada y expuesta en 2009, proponen metodologías para alcanzar el equilibrio similares a las utilizadas en Teoría de Riesgo para equilibrar riesgo y utilidad.

Tim adams y otros, en su estudio, titulado *On the fidelity versus privacy and utility trade-off of synthetic patient data*, publicado en 2025, se queja de que a pesar de las muchas medidas de calidad y modelos de generación de datos sintéticos desarrollados, la falta de estándares para la publicación de datos sintéticos conduce a incosistencias en publicaciones académicas.

Un poco más obscuro, pero no menos importante, es la existencia del trilema de la generación de datos sintéticos, que es una extensión del compromiso entre la privacidad y la utilidad, que sostiene lo paradójico que es la generación de datos sintéticos: se promete flexibilidad, privacidad y utilidad simultaneámente, pero maximizar una característica implica sacrificar al menos una de las otras dos. Se agradece a Tumult Labs hacer esto público.

1.1.3. Publicaciones corporativas

NVidia, más que un articulo académico, ha escrito publicaciones o catálogos de ventas que promueven el uso de datos sintéticos y de sus productos que permiten generarlos. Microsoft entrenó su SLM, llamado Phi-4, con datos sintéticos generados por deepseek R1 y lo hizo conocer en un informe de rendimiento presentado al público.

1.2. Nociones propedeúticas

Se entiende por nociones propedeúticas las nociones o ideas necesarias para el estudio de una disciplina o ciencia. La Estadística también las tiene. Entre estas podemos encontrar los conceptos, las metodologías que ha planteado y las limitaciones que ha encontrado al momento

de afrontar y resolver problemas. Un ejemplo son los casos especiales de la distribución beta para lidiar con la presencia de ceros o unos en los datos, cuya presencia en su forma estándar no admite.

Este capítulo expone al lector los términos considerados claves, una metodología general propuesta y las limitaciones identificadas durante el desarrollo de este proyecto. Se presentan en el orden en el cual ya fueron mencionadas.

1.2.1. Conceptos clave

Tener definiciones claras y buenas permite razonar correctamente, lo cual se manifiesta en la capacidad personal e institucional de establecer relaciones correctas entre los elementos identificados en los datos y las herramientas para analizarlos y tratarlos, que a su vez permite estudiar los diversos fenómenos que se presentan de manera óptima. A continuación se presentan las definiciones que para este trabajo se consideraron más relevantes.

- Datos Sintéticos: conjuntos de datos generados **artificialmente** por un algoritmo o modelo, a menudo de Machine Learning, que replican las propiedades estadísticas (distribución, correlaciones, tendencias,...) de un **conjunto de datos real** , pero que se espera que **contengan pocas o ninguna observación real**. Su propósito principal es proteger la **privacidad personal o corporativa** (anonimización) o aumentar el tamaño de de los datos (*data augmentation*).

- Capa o Nivel de los Datos: refiere a la **estructura jerárquica** o la **escala** en la que se organizan y analizan los datos. En un contexto de análisis complejo (como modelos multinivel o datos longitudinales), las capas representan distintos grupos de observación. Ejemplo: Estudiante individual (Nivel 1), Universidad (Nivel 2).

- Divergencia de Jensen-Shannon: mide qué tan distintas son dos distribuciones de

probabilidades, denotadas P y Q. Su fórmula para distribuciones discretas es:

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \tag{1}$$

. Con

$$M = \frac{1}{2}(P+Q)$$

y

$$D_{KL}(P\|Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right)$$

Puede tomar valores entre 0 y 1; si se aproxima a 0, las distribuciones se asemejan, pero si se aproxima 1, entoces no se asemejan.

- Divergencia de Hellinger: medida de disimilitud entre dos distribuciones de probabilidad. Su fórmula para distribuciones discretas es:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p_i}-\sqrt{q_i})^2} \tag{2}$$

Puede tomar valores entre 0 y 1; si se aproxima a 0, las distribuciones se asemejan, caso contrario si se aproxima a 1.

- Distribución aproximada: función matemática que describe el **comportamiento probabilístico** de una variable aleatoria. Una distribución aproximada es una **simplificación** elegida por el analista (ej. Normal, Binomial) para modelar la variable y facilitar el análisis o la simulación.

- Simulación: proceso de **replicar** un sistema o proceso real mediante un modelo matemático o estadístico. En este contexto, es el proceso de **generar datos sintéticos** que imiten las propiedades y el comportamiento de los datos originales.

- Datos faltantes: valores **no registrados** o **desconocidos** para una o más variables en un

conjunto de datos. Su presencia puede sesgar los resultados del análisis y es crucial identificar su patrón.

- Imputación de datos faltantes: proceso de **reemplazar** los valores faltantes con valores sustitutos (**estimados o calculados**) para crear un conjunto de datos completo que pueda ser analizado. *Métodos comunes: Imputar con la media o la imputación múltiple (MICE).*

- Parentesco o semejanza: **grado de semejanza o fidelidad estadística** entre el conjunto de **datos sintéticos** y el conjunto de **datos reales** que se está imitando. Mide qué tan bien la simulación capturó las propiedades estadísticas del original.

- Coincidencia: ocurre cuando una observación sintética es **idéntica** a un registro real en la misma posición o cuando el conjunto sintético permite **inferir, reconstruir o revelar** información sensible sobre una persona real. Este fenómeno indica un riesgo de **fuga de privacidad** o **memorización del modelo generador**. Un sistema de generación de datos sintéticos debe **minimizar las coincidencias**, garantizando que los datos simulados sean estadísticamente consistentes con los reales sin reproducir individuos específicos.

- Propensión (datos sintéticos): mide la capacidad de un clasificador para distinguir entre datos reales y sintéticos. Formalmente, sea $D = \{(x_i, y_i)\}_{i=1}^{n}$ un conjunto combinado donde $y_i = 1$ indica dato real y $y_i = 0$ dato sintético. Si un modelo estima la probabilidad $\hat{p}_i = P(y_i = 1 \mid x_i)$, la propensión idealmente cumple:

$$E[\hat{p}_i] \approx 0{,}5,$$

  Lo que indica que los datos sintéticos son estadísticamente indistinguibles de los reales. Valores cercanos a 0 o 1 indican que el modelo los separa fácilmente, lo que implica baja similitud o posible fuga de información.

- Teorema de Sklar: establece que la distribución de probabilidad conjunta de un conjunto de variables aleatorias (como la altura y el peso) puede descomponerse en dos partes

separadas: las distribuciones marginales (los patrones de cada variable de forma individual) y una función llamada cópula. La cópula es la clave, ya que es la única función que contiene toda la información sobre la estructura de dependencia o la relación entre las variables, independientemente de sus distribuciones individuales. Esto permite modelar y simular sistemas complejos, asegurando que los datos generados respeten tanto los patrones únicos de cada variable como la forma en que se influyen mutuamente.

Implicación práctica del teorema: no es posible generar datos sintéticos multivariados de calidad sin considerar su estructura de dependencia o cópula.

1.2.2. Metodología y toma de decisiones

El procedimiento a seguir en este estudio es el siguiente: I) se realiza el análisis estadística y exploratorio de los datos, II) se realiza la sintetización de los datos categóricos, III) se sintetizan los datos cuantitativos, IV) de acuerdo a las medidas de semejanza, privacidad, costo computacional y tiempo de generación, se escogen a los mejores conjuntos de datos.

En el primer paso se identifcan la dimensión de los datos (cantidad de individuos observados y covariables), los tipos de variables y la presencia de datos faltantes. Es fundamental, puesto que la correcta selección de las pruebas de semejanza depende de la correcta identificación del tipo de cada variable; las dimensiones permiten escoger los métodos de simulación, que sean adecuados para el hardware del o de los equipos qué generarán los datos; los datos faltantes, en este caso, no se imputaN, sino que se replican su frecuenia y patrones.

En el segundo paso, se generan los datos categóricos sintéticos por medio de distintos métodos, algunos basados en reglas[2] y otros no. Los métodos usados fueron: árbol de probabilidad condicional basado en reglas, cópulas gaussianas, generación tabular mediante autocodificadores variacionales y md-FAST. El paso tres fue similar al segundo, con la diferencia de que a los métodos ya probados se le añadieron elementos o probaron nuevos elementos; los nuevos métodos fueron: árbol de probabilidad condicional con reordenamiento de elementos

[2]Esto indica que un experto le indicó al algoritmo cómo simular los datos.

basados en reglas, CTGAN y CópulasGAN.

En el cuarto paso, se escogieron los datos categóricos generados considerando los resultados de las respectivas propensiones en primera instancia. En segunda instancia, se consideraron los resultados de la prueba χ^2, las distancias o divergencias de Jensen Shannon y Hellinger, el tiempo de generación o ejecución y costo computacional. Para escoger los datos cuantitativos generados se mantuvo casi todo igual, sólo se sustituyó la prueba χ^2 por la KS complemento.

1.2.3. Los límites de mi simulación son los límites de mi datos

Para terminar, es necesario considerar que los datos sintéticos son replicas de los datos reales y sus características, por ese motivo no sólo heredan sus capacidades sino también sus limitaciones. Por ejemplo: si tengo datos que permiten estimar el tiempo de supervivencia de un paciente con cáncer de pulmón, mas no su comprensión lectora, su replica heredará esta limitación. Se espera que una buena simulación herede o se acerque a la capacidad predictiva de los datos originales, que puede ser mayor o menor dependiendo del fenómeno estudiado. En ese orden de ideas, se exhorta a las personas e instituciones a reconocer las capacidades de los datos reales disponibles, de modo que hagan una crítica adecuada de los datos sintéticos y, en consecuencia, puedan proponer mejoras cuando sea necesario y aprovecharlos de la mejor manera.

2. Caso: mejores resultados de la prueba Saber Pro

2.1. Sobre los datos

Se trabajará con datos relacionados a el examen Saber Pro, el cual es “un instrumento de evaluación estandarizada para la medición externa de la calidad de la educación superior que evalúa las competencias de los estudiantes que están próximos a culminar los distintos programas profesionales universitarios” (Icfes, 2024, p.1). Se divide en cinco módulos génericos, es decir,

módulos que deben presentar todas las personas sin importar su formación profesional, además de módulos específicos en algunos casos.

Los datos que se tratan fueron recolectados entre 2016 y 2020, y comprenden una pequeña parte del total de estudiantes que presentaron el examen, quienes el ICFES consideró que obtuvieron los mejores puntajes. Este conjunto de datos fue escogido porque presentan datos mixtos, por lo cual deben simularse datos categóricos y numéricos, y su relación mutúa; además, su simpleza dilucida las dificultades que se pueden presentar cuando se generan datos sintéticos, tanto técnicas como estadísticas. El cuadro 1 resume cuántos estudiantes se presentaron y fueron considerados los mejores de sus respectivos años.

Cuadro 1: total estudiantes evaluados

Año	Evaluados	Mejores	Mejores(%)
2016	245162	11033	4.5003
2017	248855	10763	4.3250
2018	240403	10166	4.2287
2019	269431	10691	3.9679
2020	253114	12712	5.0222

Puede parecer innecesario generar datos sintéticos de una prueba que evalúa a estudiantes próximos a finalizar su formación profesional, pero hasta el 2020 se mostró información personal del estudiante en los reportes publicados, su nombre. Si se considera que además de su nombre, sus datos de residencia, formación y centro de estudio también estaban presentes, parte de su vida privada había sido expuesta, lo cual es legal y había sido acordado entre el estudiante, las instituciones de educación superior (IES) y el ICFES, por medio de un acuerdo de manejo de datos en corcondancia con la Ley 1581 de 2012, mas eso no implica que alguien externo pueda disponer de estos datos sin más.

2.1.1. Análisis descriptivo y exploratorio

Este paso es crucial, puesto que antes de generar los datos sintéticos es necesario identificar los tipos de las variables, ya que de una correcta identificación depende que se realicen las pruebas de calidad. Por ejemplo: identificar una variable tipo float como objeto, luego generar los datos sínteticos y evaluarlos con una prueba χ^2 no tiene ningùn sentido.

Las cuadros 2 y 3 resumen las dimensiones, los tipos de variables y la presencia de datos faltantes de los data sets usados para generar los datos sintéticos. Se notará que todos los datasets correspondientes al periodo comprendido entre 2016 y 2020 presentan datos faltantes mientras los posteriores a 2021 no, lo que indica una mejora en el proceso de registro de datos a partir del 2021. Los resultados obtenidos a partir del 2021 no se tratarán y simularán, pero es loable esta mejora.

En el cuadro 3 destaca la ausencia de *Percentil* si *Puntaje* está ausente, lo que revela una estrecha relación entre ambas variables. Ninguna variable categórica presenta datas faltantes. El cuadro 2 resume las dimensiones y la proporción de datos faltantes respecto al total de datos. A pesar de lo que se observa en el cuadro 3, no es ncesario imputar los datos faltantes, ya que se consideran parte del conjunto real y su presencia debe replicarse también. .

Cuadro 2: resumen de datos faltantes por año

Año	**Observaciones**	**Covariables**	**Datos**	**Datos faltantes(%)**	**¿Imputar?**
2016	11033	16	176528	1.56	Sí
2017	10763	16	172208	0.06	Sí
2018	10166	16	162656	0.49	Sí
2019	10691	16	171056	0.47	Sí
2020	12712	16	203392	0.82	Sí
2021	14222	17	241774	0.00	No
2022	11212	17	190604	0.00	No
2023	10986	17	186762	0.00	No
2024	13007	17	221119	0.00	No

Cuadro 3: Variables que presentan datos faltantes por año: 2016-2020

	Tipo	2016	2017	2018	2019	2020
Percentil Comunicación Escrita	int64	0	0	81	81	216
Puntaje Comunicación Escrita	int64	0	0	81	81	216
Percentil Razonamiento Cuantitativo	int64	331	0	0	0	0
Puntaje Razonamiento Cuantitativo	int64	331	0	0	0	0
Percentil Lectura Crítica	float64	392	21	103	127	187
Puntaje Lectura Crítica	float64	392	21	103	127	187
Percentil Competencias Ciudadanas	float64	291	14	132	92	252
Puntaje Competencias Ciudadanas	float64	291	14	132	92	252
Percentil Inglés	float64	364	19	82	101	183
Puntaje Inglés	float64	364	19	82	101	183

3. Simulación de variables categóricas

La simulación propuesta basada en reglas fue remuestrearlas a partir de un árbol de decisión basado en la probabilidad condicional de estas, las cuales se contruyen a partir de las relaciones que existen en los datos reales. Por tanto, es necesario identificar las relaciones que existen entre estas. Las relaciones identificadass se van a expresar en términos de contenencia o de pertenencia. Las relaciones identificadas fueron las siguientes:

$$\text{Departamento} \supset \text{Municipio} \supset \text{Institución} \supset \text{Programa}$$

$$\text{Programa} \in \text{Grupo de referencia}$$

Lo anterior permite construir tablas y árboles, con los cuales podemos modelar y calcular la probabilidad. En términos más formales cada probabilidad condicional se calcula de esta manera:

$$P(\mathrm{D} \cap \mathrm{M} \cap \mathrm{I} \cap \mathrm{P} \cap \mathrm{G}) = P(\mathrm{D})\, P(\mathrm{M} \mid \mathrm{D})\, P(\mathrm{I} \mid \mathrm{D} \cap \mathrm{M})\, P(\mathrm{P} \mid \mathrm{D} \cap \mathrm{M} \cap \mathrm{I})\, P(\mathrm{G} \mid \mathrm{D} \cap \mathrm{M} \cap \mathrm{I} \cap \mathrm{P})$$

Donde:

- D identifica al departamento.
- M al municipio.
- I a la institución de educación superior.
- P al programa académico.
- G al grupo de referencia.

Cada programa pertenece a un grupo exclusivamente, por lo cual $P(\mathrm{G} \mid \mathrm{P}) = 1$ si el programa pertenece a ese grupo y 0 en caso contrario. A continuación se presenta un ejemplo práctico,

basado en ejemplo 3 del cuadro 5:

$$P(Bog \cap Med \cap USTA_{Bog} \cap IngC \cap Ing) = P(Bog)\, P(Med \mid Bog) \ldots \quad = 0.50 \times 0 \times \ldots = 0.$$

La posibilidad de que Medellín sea una localidad de Bogotá y que la sede de la Universidad Santo Tomás en Bogotá esté en Medellín es 0. Una posibilidad de cero, en función de las tablas y árboles de probabilidad condicionales, le facilita al sistema detector diferenciar los datos simulados de los reales. No obstante, si la relación entre variables simuladas no difiere de una real, la dificultad del sistema para diferenciarla aumenta. Por tanto, la finalidad de simular las variable a partir de sus relaciones en los datos reales se considera una buena estrategia para simularlos. El proceso propuesto es:

I. Identificar cada variable.

II. Construir las tablas y árboles de probabilidad condicional.

III. Realizar un muestreo secuencial condicional:

    1. Selección aleatoria del Departamento de la Institución ($P(\text{Departamento})$).
    2. En función del Departamento seleccionado, elegir el Municipio aleatorio ($P(\text{Municipio} \mid \text{Departamento})$).
    3. En función del Municipio seleccionar una IES aleatoria ($P(\text{IES} \mid \text{Municipio})$).
    4. De acuerdo a la IES seleccionar un Programa Académico aleatorio ($P(\text{Programa} \mid \text{IES})$)
    5. El Grupo de Referencia se determina por el Programa Académico seleccionado ($P(\text{Grupo} \mid \text{Programa})$).

IV. Evaluar la calidad estadística de la simulación.

3.1. Pruebas de calidad estadística y elección de los datos

Recuérdese que no solo se realizó el árbol de decisión, sino que también se usaron TVAE, cópula guassiana y distribuciones marginales rapidas (MD-FAST)[3]. Estas metodologías se explican en los anexos. Se escogió el método que mejor propensión tenía, es decir aquel cuyo promedio más se acercara a 0,5.

Cuadro 4: Métricas árbol de probabilidades

Variable	**Coincidencia(%)**	χ^2	**JS**	**H**	**Autoencoder**
Grupo de Rerencia	12.47	0.9986	0.0057	0.0057	Random Forest
Nombre del Programa Académico	3.01	1.0000	0.0414	0.0434	Random Forest
Departamento de la Institución	17.77	0.9301	0.0087	0.0090	Random Forest
Municipio de la Institución	21.33	0.9991	0.0149	0.0158	Random Forest
Nombre de la Institución	3.12	1.0000	0.0270	0.0279	Random Forest

Cuadro 5: Propensiones y tiempos

Método	**Prop. 1**	**Prop. 0**	**T. de ejecución**
Árbol	0.5216	0.478	23 mins
Cópulas	0.9893	0.0227	15 mins
TVAE	0.7277	0.2834	2 horas
FAST-MD	0.1525	0.9996	7 mins

[3]Este es un método para lidiar con los elevados tiempos de generación y evaluación de los datos de laS GAN, cópula GAN y TVAE

3.2. Elección del método

El método que mejor desempeño tuvo fue el árbol de decisión, aunque supuso un mayor tiempo de ejecución. Esto se debe a que, como lo indica la tesis de Church-Turing, a mayor complejidad mayor tiempo de ejecución. En este trabajo, cabe aclarar, se recurre al tiempo para abordar lo relacionado al costo y la complejidad de computo de manera indirecta, lo cual simplifica su tratamiento.

4. Simulación de variables númericas

A diferencia de las variables categóricas, las variables numéricas presentan datos faltantes, lo que aumenta su complejidad de procesamienta, puesto que no se imputarán, al añadir un estado adicional. Esto se evidenció en los tiempos de ejecución, como se verá más adelante.

Los resutados de los múltiples intentos que se les expondrán muestran que es un error gravísimo separar los datos numéricos del contexto del que se originan, que en este caso se puede conocer por las variables categóricas, e ignorar sus relaciones internas como la correlación, por ejemplo, la cual es fundamental para diferenciar una buena simulación de una mala, ya que esta permite conocer si se preservaron o no las relaciones entre las variables.

Se usaron 6 métodos para generar los datos sintéticos. El primero consistió en la simulación de cada variable con su respectiva distribución marginal y parámetros. El segundo en modelar los datos sintéticos con cópulas gaussianas [4]. El tercero se hizo siguiendo una cópula GAN[5]. De cuarto y quinto métodos se propusieron una CTGAN y un MD-FAST, cuyas explicaciones intuitivas encontrarán en los anexos. El sexto fue un árbol de decisión combinado con un reordamiento de las variables cuantitativas, todo dirigido bajo la figura de un *experto*.

[4] Si desea conocer el procedimiento seguido, puede revisar el anexo I.
[5] Si desea conocer el procedimiento seguido, puede revisar el anexo II.

4.1. Método I: simulación individual por variable

Los resultados de este método se presentan a continuación. La primera impresión, según lo indiciado en el cuadro 7, indica que son similares a sus contrapartes reales, pero el cuadro 8 muestra que sus relaciones de dependecia no fueron captadas. Esto se prevee desde el **teorema de Sklar**, que establece que cualquier distribución multivariada se puede descomponer en dos partes: las distribuciones marginales de cada variable y cópula o estructura de dependencia. La corrección, por tanto, parece ser la inclusión de la cópula cuando se simulan los datos. El cuadro resume las distribuciones marginales de las variables cuantitaivas.

Cuadro 6: Análisis de Distribuciones Marginales y Colas

Variable	**Distribución**	**Mejor Ajuste (Score)**	**Asimetría**	**Colas Pesadas**
CE	Dweibull	0.00213	Asimétrica (Flexible)	Sí (si $k < 1$)
RC	Genextreme	0.00335	Asimétrica (Flexible)	Sí
LC	Dweibull	0.00223	Asimétrica (Flexible)	Sí (si $k < 1$)
CC	Normal	0.00163	Simétrica	No (colas ligeras)
ING	t	0.00126	Simétrica	**Sí**)

Nota: se realizaron pruebas con cópulas guassianas y cópulas GAN, sin incuir las variables cualitativas y ambas no difirieron del resultado del método I.

Cuadro 7: Fidelidad de las Distribuciones Marginales

Variable	**KS Complement**	**Jensen Shannon**	**Hellinger**
PUNTAJE LECTURA CRÍTICA	0,9587	0,0977	0,1155
PUNTAJE COMUNICACIÓN ESCRITA	0,9626	0,0957	0,1130
PUNTAJE RAZONAMIENTO CUANTITATIVO	0,9495	0,0422	0,0422
PUNTAJE COMPETENCIAS CIUDADANAS	0,9732	0,0692	0,0790
PUNTAJE INGLÉS	0,9443	0,1163	0,1382
PUNTAJE GLOBAL	0,9804	0,0073	0,0073

Cuadro 8: comparación correlación datos sintéticos contra datos sintéticos

$\rho_{i,j}$	ρ_R	ρ_S
$\rho_{LC,PG}$	0,6177	0,0048
$\rho_{CC,PG}$	0,5614	0,0017
$\rho_{ING,PG}$	0,5566	0,0097
$\rho_{RC,PG}$	0,5140	0,0021
$\rho_{CE,PG}$	0,4943	0,0055
$\rho_{LC,CC}$	0,2629	0,0036
$\rho_{LC,ING}$	0,1974	−0,0032
$\rho_{CC,ING}$	0,1888	0,0099
$\rho_{CE,CC}$	0,1587	−0,0017
$\rho_{CE,ING}$	0,1297	−0,0020
$\rho_{RC,CC}$	0,1060	0,0021
$\rho_{LC,CE}$	0,1118	0,0087
$\rho_{LC,RC}$	0,0978	0,0006
$\rho_{RC,ING}$	0,0394	−0,0019
$\rho_{RC,CE}$	0,0063	−0,0011

4.2. Propensión de los métodos usados

Se escogió el método en función de la menor detección de datos sintéticos. En caso de que hubieran resultados similares, se consideró el tiempo de ejecución como criterio de desempate, esto se decidió debido a que si dos procesos llevan al mismo resultado, pero uno es más rápido que el otro, se prefiere el más rápido. La propensión, aquí se presenta como evasión y detección, que indican que porcentaje de los datos sintéticos fue detectado y cuál no.

Cuadro 9: Propensiones y tiempos por modelo

	Evasión	**Detección**	**T. de ejecución**	**Autoenconder**
Árbol+Shuffle	0	1	3 horas	Reg. logística
Cópulas	0.3114	0.6886	3 mins.	Reg. logística
Cópulas GAN	0.3923	0.6077	5 horas	Reg. logística
FAST-MD	-	-	P. finalizados sin éxito	Random Forest
CTGan	-	-	P. finalizados sin éxito	Reg. logística

4.3. Método seleccionado: cópulas gaussianas

El método escogido fue la cópula gaussiana, ya que presentó una detección similar a las cópulas GAN, pero su tiempo de ejecución fue menor. Lo anterior no implica que haya sido una simulación óptima, pero *lo perfecto es enemigo de lo bueno* y en ese orden de ideas, por lo menos cumple con enmascarar y un tiempo viable de ejecución en la vida real. Si se tratara de sólo enmascarar, la Cópula GAN es muy superior.

El cierre de este capítulo es agridulce en cuanto a que las simulaciones no fueron las adecuadas, pero un éxito ya que reveló un problema bastante serio: capturar la totalidad de la estructura de dependencia, conforme aumentan las variables que interactuan entre sí. Las imágenes generadas por Matejka y Fitzmaurice (citados por Juan R., 2023), confirman que es posible tener semejanza estadística en medidas d1e tendencia central o de dispersión, incluso de correlación lineal, y áun fallar en capturar la estructura de dependencia.

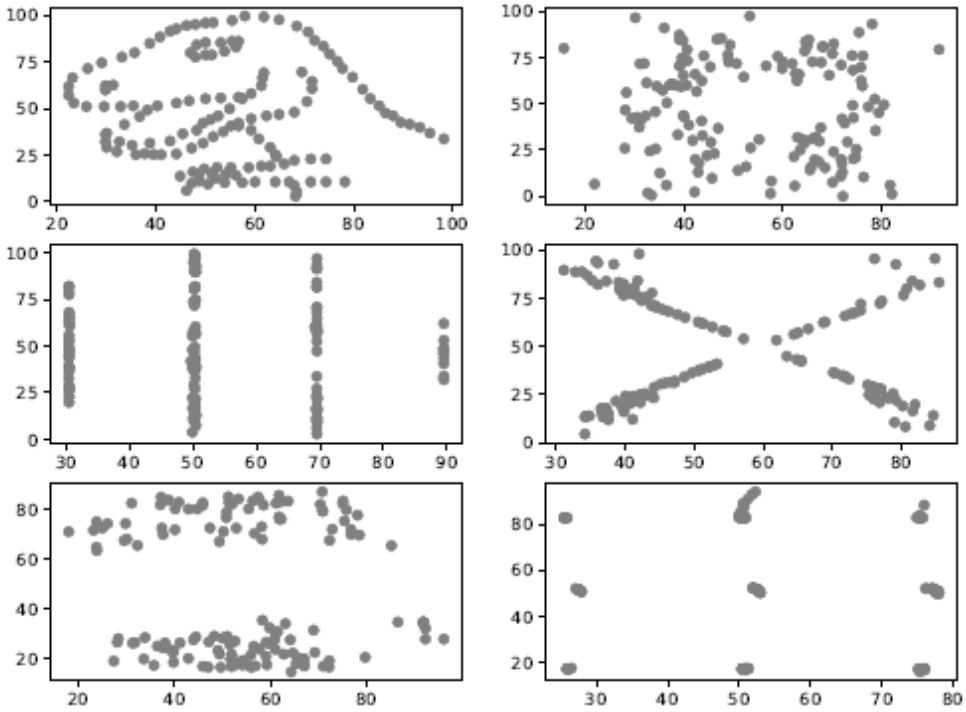

Figura 1: misma media, misma varianza, misma correlación lineal, pero distinta forma.

Cuadro 10: Resultados de similitud de correlación

Índice	$\rho_{i,j}$	ρ_R	ρ_s
0	$\rho_{LC,CE}$	0.113360	0.129931
5	$\rho_{CE,CC}$	0.158885	0.174483
8	$\rho_{RC,ING}$	0.040153	0.051539
7	$\rho_{RC,CC}$	0.107981	0.119295
1	$\rho_{LC,RC}$	0.098730	0.104467
6	$\rho_{CE,ING}$	0.130028	0.135383
4	$\rho_{CE,RC}$	0.007076	0.003787
3	$\rho_{CE,CC}$	0.199166	0.200972
9	$\rho_{CC,ING}$	0.191291	0.190829
2	$\rho_{LC,CC}$	0.264706	0.264295

5. Discusión final

Los datos sintéticos no son lo que los anglosajones llaman una *silver bullet*[6], sin embargo, como toda herramienta, tiene capacidades y limitaciones, fortalezas y debilidades. En esta última sección se expondrán algunas debilidades de estos y estrategias propuestas para superarlas.

5.1. Debilidades

En este trabajo se compararon distintos métodos para simular variables categóricas y numéricas. Los datos sintéticos categóricos en comparación con los cuantitativos parecieron tener una calidad mayor y no presentar problemas, mas no es cierto, puesto que la simulación categórica tiene un problema bastante serio: presenta pérdida de categórias.

Al publicar datos sintéticos categóricos, como ocurre con nuestro caso de estudio, se corre el riesgo de invisibilizar a grupos en riesgo. Por ejemplo: dada una enfermedad, es posible borrar a un grupo de pacientes que padecen una variante rara o muy poco probable de contraer. Este fenómeno es problemático, ya que falla en ser útil no en el sentido estadístic, sino en el médico, es decir, en asegurar la protección de ciertos pacientes. Es decir, la aplicación de estos

La pérdida de categorías puede ser causal para rechazar un estudio, ya que se violan principios éticos de inclusión y representación de minorías y población vulnerable. Por ejemplo: el DANE se rige por un sistema ético estadístico llamado SETE, el cual investiga y rechaza cuando se borra una minoría o población vulnerable, lo cual complica el uso de datos sintéticos. Ya se conoce la dificultad que se enfrenta cuando se simulan datos cuantitativos, la captura de la cópula es más difícil conforme aumentan las interacciones entre variables. ¿Qué soluciones han sido propuestas?

[6] Solución simplista y en apariencia mágica que puede resolver problemas complejos.

Cuadro 11: Resumen comparativo de categorías unicas y moda

Variable	Cat. Sintéticas	Cat. Reales	Cat. Perdidas	Moda (Tabla 1)	Moda (Tabla 2)
Depto. institución	27	27	0	Bogotá	Bogotá
Municipio institución	70	80	**10**	Bogotá, D. C.	Bogotá, D. C.
Institución	252	266	**14**	Los Andes	Los Andes
Programa académico	666	710	**44**	Derecho	Derecho
Grupo de referencia	23	23	0	Ingeniería	Ingeniería

5.2. Soluciones propuestas

5.2.1. Categorías que desaparecen: recalibración de los pesos

No en todas las disciplinas es igual de peligroso que desaparezcan categorías, pero en la medicina ciertamente lo es. Ellos han propuesto recalibrar los pesos con el fin de que los eventos con muy baja probabilidad no sean borrados en las simulaciones.

5.2.2. Capturar la cópula en su totalidad

Capturar y replicar la cópula en su totalidad es teóricamente posible, pero en la realidad implica poseer una explicación perfecta del fenómeno, lo cual es imposible. La solución es recordar lo dicho por George Edward Pelham Box: «Todos los modelos son erróneos, pero algunos son útiles». Si una simulación es los suficientemente buena para enmascarar debe así reconocerse, no será perfecta, pero si funciona, funciona, y eso basta. Conforme mejoren las técnicas de generación, también los datos sintéticos generados.

6. Anexos

6.1. Cópulas gaussianas (normales)

La cópula gaussiana es una función estadística crucial que modela la estructura de dependencia entre variables aleatorias, basándose en la robustez de la distribución normal multivariada.

6.1.1. Definición formal

Una cópula gaussiana C con matriz de correlación Σ se define como la función de distribución acumulada (CDF) de un vector Gaussiano estandarizado. Permite relacionar una función de distribución conjunta H con sus distribuciones marginales $F_1, ..., F_d$.

$$C(u_1, ..., u_d; \Sigma) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d))$$

Donde:

- $u_i \in [0,1]$ es la probabilidad marginal transformada, $u_i = F_i(x_i)$.
- Φ^{-1} es la función cuantil (inversa de la CDF) de la distribución normal estándar $\mathcal{N}(0,1)$.
- Φ_Σ es la CDF de la distribución normal multivariada con matriz de correlación Σ.

6.1.2. Características clave

- **Soporte:** La cópula opera sobre el cubo unitario $[0,1]^d$.
- **Simetría de Cola:** Presenta una dependencia simétrica. La dependencia en la cola superior (valores extremos altos) es idéntica a la dependencia en la cola inferior (valores extremos bajos).

- **Limitación:** Debido a su simetría, tiende a subestimar la dependencia en escenarios donde las caídas son más correlacionadas que las subidas (ej., mercados financieros).

6.1.3. Procedimiento de aplicación (modelado y simulación)

La aplicación de la cópula gaussiana para modelar la dependencia entre un vector de variables $\mathbf{X} = (X_1, ..., X_d)$ sigue tres pasos esenciales:

Paso I: Transformación a Uniforme (Marginales)

Se transforma cada variable X_i a una variable uniforme U_i en $[0,1]$ mediante su función de distribución acumulada empírica o paramétrica F_i:

$$U_i = F_i(X_i)$$

6.1.3.1 Paso II: transformación a normal estándar

Las variables uniformes U_i se transforman a variables normales estándar Z_i mediante la función cuantil Φ^{-1}:

$$Z_i = \Phi^{-1}(U_i)$$

6.1.3.2 Paso III: Estimación del Parámetro de Dependencia

Se utiliza la matriz de correlación Σ de las variables transformadas $\mathbf{Z} = (Z_1, ..., Z_d)$ como el parámetro de la cópula.

$$\Sigma = \text{Corr}(\mathbf{Z})$$

Una vez determinada Σ, la cópula está completamente especificada y puede ser utilizada para calcular probabilidades conjuntas o generar nuevos datos sintéticos $\mathbf{X}^*$ mediante el proceso inverso de des-transformación.

6.2. Cópula GAN

Las Cópulas GAN (Copula-based Generative Adversarial Networks) no son un tipo de cópula en sí mismas (como la Gaussiana o la Gumbel), sino una metodología de modelado y estimación no paramétrica que utiliza la potencia de las GANs para resolver los desafíos de las cópulas en alta dimensión

6.2.1. Fase I: modelado marginal y transformación

Esta fase se centra en aislar la forma de la distribución de cada variable X_i y transformarla a un espacio uniforme.

6.2.1.1 Paso A: transformada integral de probabilidad (PIT)

Las distribuciones originales de los datos reales $\mathbf{X} = (X_1, \ldots, X_d)$ se transforman individualmente a una distribución uniforme estándar.

1. Para cada variable X_i, se ajusta su Función de Distribución Acumulada (CDF) empírica o paramétrica, F_i.

2. Se aplica la **Transformada Integral de Probabilidad** (PIT) para obtener una variable uniforme $U_i \sim \mathscr{U}(0,1)$.

$$U_i = F_i(X_i)$$

Razón: Los datos transformados $\mathbf{U} = (U_1, \ldots, U_d)$ retienen la estructura de dependencia original, pero con marginales uniformes, lo que simplifica el modelado de la cópula.

6.2.2. Fase II: modelado de dependencia (la cópula)

La cópula C es una función que une las distribuciones marginales F_i para formar la distribución multivariada F.

6.2.2.1 Paso B: ajuste de la cópula

Según el **Teorema de Sklar**, existe una cópula C tal que:

$$F(x_1,\ldots,x_d) = C(F_1(x_1),\ldots,F_d(x_d)) = C(u_1,\ldots,u_d)$$

El modelo ajusta una cópula (usualmente una **cópula gaussiana**) al vector **U**.

6.2.2.2 Paso C: transformación a espacio Normal

Para facilitar el entrenamiento de la GAN, el vector uniforme **U** se transforma al **espacio Normal estándar** ($\mathbf{Z} \sim \mathcal{N}(0,1)$) utilizando la inversa de la CDF de la Normal Estándar (Φ^{-1}):

$$\mathbf{Z} = (\Phi^{-1}(U_1),\ldots,\Phi^{-1}(U_d))$$

Razón: Los datos en el espacio normal son ideales para el entrenamiento de GANs, ya que las distorsiones se controlan mejor.

6.2.3. Fase III: generación y reconstrucción GAN

Se entrena un par Generador-Discriminador para aprender a sintetizar datos $\mathbf{Z}_{\text{sim}}$ que repliquen la estructura de $\mathbf{Z}_{\text{real}}$.

6.2.4. Paso D: entrenamiento de la GAN

1. El **Generador** G recibe un vector de ruido aleatorio z_{ruido} y produce datos sintéticos $\mathbf{Z}_{\text{sim}}$.
2. El **Discriminador** D intenta distinguir entre $\mathbf{Z}_{\text{sim}}$ y $\mathbf{Z}_{\text{real}}$.
3. El entrenamiento se rige por la función de pérdida adversarial (minimización/maximización) que obliga a G a capturar la **estructura de correlación** presente en $\mathbf{Z}_{\text{real}}$.

Paso E: transformación inversa Una vez que el Generador produce $\mathbf{Z}_{\text{sim}}$ de alta calidad, se revierten las transformaciones para obtener los datos finales $\mathbf{X}_{\text{sim}}$:

1. De Normal a Uniforme: $\mathbf{U}_{\text{sim}} = \Phi(\mathbf{Z}_{\text{sim}})$.
2. De Uniforme a Marginal Original: $\mathbf{X}_{\text{sim}} = (F_1^{-1}(U_{\text{sim},1}), \ldots, F_d^{-1}(U_{\text{sim},d}))$.

6.2.5. Resultado

El DataFrame $\mathbf{X}_{\text{sim}}$ generado por CopulaGAN satisface dos criterios clave simultáneamente:

- **Fidelidad Marginal:** Cada columna $\mathbf{X}_{\text{sim},i}$ sigue de cerca la distribución original F_i. (Alto Score de *Column Shapes*).
- **Fidelidad de Dependencia:** Las correlaciones $\rho(\mathbf{X}_{\text{sim},i}, \mathbf{X}_{\text{sim},j})$ son fieles a las correlaciones reales $\rho(\mathbf{X}_{\text{real},i}, \mathbf{X}_{\text{real},j})$. (Alto Score de *Column Pair Trends*).

6.3. TVAEs

Las *Tabular Variational Autoencoders* (TVAE) son modelos generativos basados en redes neuronales diseñados específicamente para datos tabulares. Su objetivo es aprender la distribución conjunta de un conjunto de variables a partir de datos reales y, a partir de este aprendizaje, generar nuevas observaciones sintéticas con propiedades estadísticas similares.

El modelo se compone de dos partes principales: un codificador y un decodificador. El codificador transforma cada observación original en una representación latente de menor dimensión, mientras que el decodificador reconstruye los datos a partir de dicha representación. Durante el entrenamiento, el modelo aprende a equilibrar la calidad de la reconstrucción con la regularización del espacio latente, lo que permite generar datos nuevos y no copiados.

Una vez entrenado, la generación de datos sintéticos se realiza muestreando valores aleatorios desde el espacio latente aprendido y pasándolos por el decodificador. De este modo, las

observaciones sintéticas conservan tanto las distribuciones marginales como las relaciones de dependencia entre las variables originales.

6.4. CTGAN

El *Conditional Tabular Generative Adversarial Network* (CTGAN) es un modelo generativo basado en redes generativas adversarias diseñado para datos tabulares. A diferencia de los GAN tradicionales, CTGAN incorpora mecanismos específicos para manejar variables categóricas y distribuciones numéricas complejas, comunes en este tipo de datos.

El modelo consta de dos redes neuronales: un generador y un discriminador, que compiten entre sí durante el entrenamiento. El generador produce observaciones sintéticas a partir de ruido aleatorio y condiciones impuestas sobre ciertas variables, mientras que el discriminador intenta distinguir entre datos reales y sintéticos. Este esquema adversarial permite al generador aprender la distribución conjunta de los datos originales.

La generación condicional es un elemento clave de CTGAN, ya que permite equilibrar la representación de categorías poco frecuentes y mejorar la calidad de las dependencias aprendidas. Una vez entrenado, el modelo genera datos sintéticos muestreando ruido y condiciones, produciendo observaciones que preservan tanto las distribuciones marginales como las relaciones de dependencia entre las variables.

6.5. FAST-MD

El método FAST–MD (*FAST Marginal Distributions*) es una estrategia heurística para la generación de datos sintéticos que se basa en la preservación explícita de las distribuciones marginales de cada variable y en la incorporación de dependencias simples mediante reordenamiento. A diferencia de modelos generativos complejos, FAST-MD prioriza la eficiencia computacional y la estabilidad, utilizándose principalmente como un método base (*baseline*) para comparación.

El procedimiento de FAST–MD consta de las siguientes etapas:

1. **Estimación de distribuciones marginales.** Para cada variable se estima su distribución marginal a partir de los datos reales, utilizando frecuencias relativas para variables categóricas y distribuciones empíricas o paramétricas para variables numéricas.

2. **Muestreo independiente.** A partir de las distribuciones marginales estimadas se generan nuevas observaciones de manera independiente para cada variable.

3. **Incorporación de dependencia básica.** Con el fin de evitar la independencia total entre variables, se aplica un reordenamiento por rangos (*rank shuffling*) que permite recuperar correlaciones globales de forma aproximada.

4. **Obtención del conjunto sintético.** El conjunto de datos resultante conserva adecuadamente las distribuciones marginales y dependencias simples, manteniendo un bajo costo computacional.

FAST–MD no busca modelar dependencias complejas ni interacciones de alto orden, sino ofrecer una alternativa rápida y reproducible para la generación de datos sintéticos, especialmente útil en contextos de grandes volúmenes de datos o recursos computacionales limitados.

6.6. Ejemplos de cálculo de divergencias de Jensen y Shannon

Consideramos dos distribuciones de probabilidad discretas P y Q sobre tres resultados ($i = 1, 2, 3$):

Cuadro 12: Distribuciones de Probabilidad P y Q

Resultado (i)	Distribución P (P_i)	Distribución Q (Q_i)
1	0,5	0,3
2	0,3	0,5
3	0,2	0,2

6.6.1. Pasos del Cálculo

6.6.2. Paso 1: Cálculo de la Distribución Mezcla (M)

$$M_i = \frac{1}{2}(P_i + Q_i)$$

Cuadro 13: Cálculo de la Distribución M

i	P_i	Q_i	$M_i = (P_i + Q_i)/2$
1	0,5	0,3	0,40
2	0,3	0,5	0,40
3	0,2	0,2	0,20
Suma	1.0	1.0	1.00

6.6.3. Paso 2: Cálculo de $\mathrm{D}_{KL}(P\|M)$

$$\mathrm{D}_{KL}(P\|M) = \sum_i P_i \log_2 \left(\frac{P_i}{M_i}\right)$$

$$\begin{aligned}
\mathrm{D}_{KL}(P\|M) &= 0{,}5 \log_2 \left(\frac{0{,}5}{0{,}4}\right) + 0{,}3 \log_2 \left(\frac{0{,}3}{0{,}4}\right) + 0{,}2 \log_2 \left(\frac{0{,}2}{0{,}2}\right) \\
&\approx 0{,}5(0{,}3219) + 0{,}3(-0{,}4150) + 0{,}2(0{,}0000) \\
&\approx 0{,}16095 - 0{,}12450 + 0{,}00000 \\
&\approx 0{,}03645
\end{aligned}$$

6.6.4. Paso 3: Cálculo de $\mathrm{D}_{KL}(Q\|M)$

$$\mathrm{D}_{KL}(Q\|M) = \sum_i Q_i \log_2 \left(\frac{Q_i}{M_i}\right)$$

$$\begin{aligned}
\mathrm{D}_{KL}(Q\|M) &= 0{,}3\log_2\left(\frac{0{,}3}{0{,}4}\right) + 0{,}5\log_2\left(\frac{0{,}5}{0{,}4}\right) + 0{,}2\log_2\left(\frac{0{,}2}{0{,}2}\right) \\
&\approx 0{,}3(-0{,}4150) + 0{,}5(0{,}3219) + 0{,}2(0{,}0000) \\
&\approx -0{,}12450 + 0{,}16095 + 0{,}00000 \\
&\approx 0{,}03645
\end{aligned}$$

6.6.5. Paso 4: Cálculo de la JSD$(P\|Q)$

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}\mathrm{D}_{KL}(P\|M) + \frac{1}{2}\mathrm{D}_{KL}(Q\|M)$$

$$\begin{aligned}
\mathrm{JSD}(P\|Q) &\approx \frac{1}{2}(0{,}03645) + \frac{1}{2}(0{,}03645) \\
&\approx 0{,}03645
\end{aligned}$$

6.6.6. Conclusión

La Divergencia de Jensen-Shannon entre las distribuciones P y Q es aproximadamente 0,03645. Un valor cercano a cero indica una alta similitud entre las dos distribuciones.

6.7. Distancia de Hellinger (H)

$$H(P,Q) = \sqrt{1 - \sum_i \sqrt{P_i Q_i}} = \sqrt{1 - \mathrm{Similitud}(P,Q)} \tag{3}$$

7. Datos de Ejemplo

Utilizamos las mismas distribuciones de probabilidad P y Q sobre tres resultados ($i = 1, 2, 3$):

Cuadro 14: Distribuciones de Probabilidad P y Q

Resultado (i)	Distribución P (P_i)	Distribución Q (Q_i)
1	0,5	0,3
2	0,3	0,5
3	0,2	0,2

7.0.1. Pasos del Cálculo

7.0.1.1 Paso 1: Cálculo de los Productos de las Raíces ($\sqrt{P_i Q_i}$)

Cuadro 15: Cálculo de los Productos Intermedios

i	P_i	Q_i	$\sqrt{P_i Q_i}$	Valor Aproximado
1	0,5	0,3	$\sqrt{0,15}$	$\approx 0,38730$
2	0,3	0,5	$\sqrt{0,15}$	$\approx 0,38730$
3	0,2	0,2	$\sqrt{0,04}$	0,20000

7.0.1.2 Paso 2: cálculo de la Afinidad de Bhattacharyya (Similitud)

$$\text{Similitud}(P, Q) = \sum_i \sqrt{P_i Q_i}$$

$$\text{Similitud}(P, Q) \approx 0,38730 + 0,38730 + 0,20000$$

$$\approx 0,97460$$

7.0.1.3 Paso 3: cálculo de la Distancia de Hellinger (H)

$$H(P,Q) = \sqrt{1 - \text{Similitud}(P,Q)}$$

$$H(P,Q) \approx \sqrt{1 - 0{,}97460}$$

$$H(P,Q) \approx \sqrt{0{,}02540}$$

$$H(P,Q) \approx 0{,}15937$$

7.0.2. Conclusión

La Distancia de Hellinger entre las distribuciones P y Q es aproximadamente 0,15937. Dado que este valor es cercano a cero, indica una alta similitud entre las dos distribuciones de probabilidad.

Referencias

[1] Abowd, J., & Lane, J. (2003). Synthetic data and confidentiality protection. https://www2.census.gov/ces/tp/tp-2003-10.pdf

[2] APX. (2025). Propensity score evaluation. *Evaluating Synthetic Data Quality*. https://apxml.com/courses/evaluating-synthetic-data-quality/chapter-2-advanced-statistical-fidelity/propensity-score-evaluation

[3] Colombi, K. (2025). How to generate synthetic data: A comprehensive guide. *Tonic.ai*. https://www.tonic.ai/guides/how-to-generate-synthetic-data-a-comprehensive-guide

[4] DataCebo. (2025). sdmetrics. https://docs.sdv.dev/sdmetrics

[5] Desfontaines, D. (2023). The fundamental trilemma of synthetic data generation. https://www.tmlt.io/resources/fundamental-trilemma-synthetic-data-generation

[6] El Namaki, M. S. S. (2025). AI data syndrome: Could synthetic data lead to hallucinative AI analytics outcomes? https://ijeber.com/uploads2025/ijeber$_0$5$_3$25.pdf

[7] Farhadyar, K., Bonofiglio, F., Hackenberg, M., Behrens, M., Zöller, D., & Binder, H. (2024). Combining propensity score methods with variational autoencoders for generating synthetic data in presence of latent sub-groups. *BMC Medical Research Methodology, 24*, Article 198. https://doi.org/10.1186/s12874-024-02327-x

[8] Genest, C., & Favre, A. (1986). Everything you always wanted to know about copula modeling but were afraid to ask. https://www.uni-muenster.de/Physik.TP/ lemm/seminarSS08/JHE-2007.pdf

[9] Hamming, R. (1986). You and your research. https://docs.sdv.dev/sdmetrics

[10] Jordan, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2024). Synthetic data: What, why and how? https://www.tmlt.io/resources/fundamental-trilemma-synthetic-data-generation

[11] Kooistra, E. (2024). How do I know that the synthetic data is of the right quality for my use case? https://bluegen.ai/how-do-i-know-that-the-synthetic-data-is-of-the-right-quality-for-my-use-case/

[12] Locowic, L., Monteverdi, A., & Mendoza, E. (2024). Synthetic data generation from real data sources using Monte Carlo tree search and large language models. https://www.authorea.com/users/832345/articles/1225643

[13] Mohapatra, S., Zong, J., Kerschbaum, F., & He, X. (2022). Differentially private data generation with missing data. *Proceedings of the VLDB Endowment*. https://www.vldb.org/pvldb/vol17/p2022-mohapatra.pdf

[14] Office of the Privacy Commissioner of Canada. (2022). When what is old is new again – The reality of synthetic data. *Privacy Tech-Know Blog*. https://www.priv.gc.ca/en/blog/20221012

[15] Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Intelligent Systems and Applications in Engineering, 11*(3), 100177. https://www.sciencedirect.com/science/article/pii/S2667096823000241

[16] Ravn, L. (2024). The overlooked politics of synthetic data performance metrics. https://www2.census.gov/ces/tp/tp-2003-10.pdf

[17] Restrepo Lopera, J. P. (2023). Nonparametric generation of synthetic data using copulas. https://repository.eafit.edu.co/entities/publication/e87d24db-ff18-44e6-b8b3-8eb43b1870fa

[18] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*.

https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

[19] Shannon, C. E. (1948). A mathematical theory of information. *Bell System Technical Journal, 27*(3), 379–423.
https://people.math.harvard.edu/ ctm/home/text/others/shannon/entropy/entropy.pdf

[20] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(3), 663–688.
https://academic.oup.com/jrsssa/article/181/3/663/7072005

[21] UNESCO. (2025). Ethics of artificial intelligence.
https://www.unesco.org/en/artificial-intelligence/recommendation-ethics