02

CASO DE ESTUDIO

SOBRE SIMULACIÓN DE DATOS PARA INVESTIGACIONES ACADÉMICAS MEDIANTE INTELIGENCIA ARTIFICIAL GENERATIVA Y GOOGLE COLAB

Fecha de presentación: enero, 2025 **Fecha de aceptación:** marzo, 2025 **Fecha de publicación:** abril, 2025

CASO DE ESTUDIO

SOBRE SIMULACIÓN DE DATOS PARA INVESTIGACIONES ACADÉMICAS MEDIANTE INTELIGENCIA ARTIFICIAL GENERATIVA Y GOOGLE COLAB

CASE STUDY ON DATA SIMULATION FOR ACADEMIC RESEARCH USING GENERATIVE ARTIFICIAL INTELLIGENCE AND GOOGLE COLAB

Miguel Ángel Fernández-Marín[1]
E-mail: miguelangelferssc@gmail.com
ORCID: https://orcid.org/0000-0002-6132-539X
Jordy Rafael Montero-Murillo[1]
E-mail: jordymontero01@gmail.com
ORCID: https://orcid.org/0009-0004-8158-0672
Débora González-Tolmo[2]
E-mail: dtolmo1986@gmail.com
ORCID: https://orcid.org/0000-0002-8890-130X
[1]Universidad Metropolitana. Ecuador.
[2]Empresa de software en Quito Netby. Ecuador.

Cita sugerida (APA, séptima edición)

Fernández-Marín, M. Á., Montero-Murillo, J. R., & González-Tolmo, D. (2025). Caso de estudio sobre simulación de datos para investigaciones académicas mediante Inteligencia Artificial Generativa y Google Colab. *Revista Mexicana de Investigación e Intervención Educativa, 4(S1), 18-26.*

RESUMEN

Se presenta un caso de estudio sobre simulación de datos para la investigación académica, donde con Inteligencia Artificial Generativa (IAG) y el modelo ChatGPT se logra obtener un conjunto de datos simulados con los prompts adecuados. Para la validación de datos se utilizó Google Colab para Python. La metodología empleada se enfoca en comparar dos conjuntos de datos, uno que es real y otro que no, pero construido bajo parámetros de los datos primarios como la media, desviación estándar y cantidad de datos. Se utilizaron librerías específicas de Paython como numpy y ttest_ind para el análisis de la estadística como T-Student, y otras como matplotlib y seaborn para gráficos de densidad. Los resultados arrojados contrastan que los datos simulados guardaban estrecha relación con los datos reales. Se demostró que no fue significativo las diferencias estadísticas demostrando la utilidad de la metodología empleada.

Palabras clave:

Datos simulados, inteligencia artificial generativa, validación de datos, Python, ChatGPT, prompts, t de Student.

ABSTRACT

A case study on data simulation for academic research is presented, where with Generative Artificial Intelligence (IAG) and the ChatGPT model it is possible to obtain a set of simulated data with the appropriate prompts. Google Colab for Python was used for data validation. The methodology used focuses on comparing two sets of data, one that is real and another that is not, but constructed under parameters of the primary data such as the mean, standard deviation and amount of data. Specific Paython libraries such as numpy and ttest_ind were used for the analysis of statistics such as T-Student, and others such as matplotlib and seaborn for density plots. The results show that the simulated data was closely related to the real data. It was shown that the statistical differences were not significant, demonstrating the usefulness of the methodology used.

Keywords:

Simulated data, generative artificial intelligence, data validation, Python, ChatGPT, prompts, t-Student.

INTRODUCCIÓN

En la investigación científica, es requerido la recopilación de datos de forma adecuada, para un análisis efectivo de los mismo. Esto constituye la base para poder aplicar técnicas estadísticas con el propósito de validar hipótesis, desarrollar teorías y generar nuevo conocimiento. La fiabilidad de los resultados es determinada por la confiabilidad de los datos, teniendo en cuenta su precisión y calidad.

En acuerdo con Machuca Martínez (2020), los datos de investigación son componentes esenciales que sustentan el avance científico, garantizan la calidad y reproducibilidad de los estudios, y fomentan una cultura de apertura y colaboración en la comunidad académica. Actualmente, cobran importancia los repositorios de datos estables para el desarrollo de nuevas investigaciones con un fin determinado, en consecuencia Marín Arraiza et al. (2019), definen que los datos se han convertido en la base de la infraestructura de la ciencia.

Esto tiene que ver, con la posición de compartir conocimiento científico, donde los investigadores colaboran entre si proporcionando el bien más valioso, los datos, por eso velar por su integridad es importante. Además, con la participación de especialistas, influye en que se validen y perfeccionen técnicas para lograr la adecuada estructura de los datos para su consumo y análisis, teniendo en cuenta técnicas de curado de datos. Además, el acceso a los mismo es importante, aunque muchas veces no es posible por diferentes razones, por lo que la comunidad científica debe asumir estrategias con diferentes recursos tecnológico para la validación de sus experimentos. Todo esto y en acuerdo con Morillo Moreno (2024), ha posibilitado la verificación de resultados dentro de la comunidad científica, que los análisis se hagan más profundos y se logren con esto mejores interpretaciones en el proceso científico.

Hoy en día se han incrementado el interés por formalizar repositorios internacionales, con alto porcentaje de confiabilidad, donde los investigadores comparten datos ya probados, que son muy utilizados por la comunidad científica. En acuerdo con López (2024), esto impacta positivamente en la transparencia de los resultados y en la reutilización de datos para nuevos contextos de estudios futuros. En consecuencia, facilita la manipulación de los recursos desde repositorio en conexión en línea, evitando la duplicación de esfuerzo. Con ello se aumenta la visibilidad y el impacto de la investigación.

Pero no todo es tan fácil para el acceso a conjunto de datos fiables, ya que existen inconvenientes para obtener los mismo. Una muestra de ellos lo destacan Kwok et al. (2022), cuando en su trabajo declaran que altos costos y el tiempo prolongado necesarios para la recolección de datos en investigaciones de salud pueden retrasar la obtención de resultados y limitar la viabilidad de estudios a gran escala. También Garrido Elustondo et al. (2012); Gordon (2020); y Vilches (2024), destacan que las restricciones éticas que surgen al realizar investigaciones con poblaciones vulnerables, enfatizando la necesidad de protocolos que protejan a los participantes y cómo estas restricciones pueden limitar el acceso a datos esenciales.

Conservar la ética es importante a la hora del consumo de datos para investigaciones, pues dependiendo de la fuente, pueden ser datos sensibles. Tener en cuenta siempre su legalidad a la hora del consumo de estos es crucial, dado que la comunidad científica no sólo se preocupa por su fiabilidad sino también por su acceso restringido, en dependencia de la necesidad de no ser público apegándose a las leyes de protección de datos. Por lo que el surgimiento de nuevas investigaciones constituye un reto, al ser necesario conciliar un equilibrio entre la relevancia del dato y proteger los derechos a ellos. Nuevas técnicas de datos simulados emergen con el fin de ser utilizados en investigaciones. Este nuevo contexto, relaciona tecnologías emergentes con herramientas tradicionales para fortalecer conjuntos de datos para la validación de investigaciones, teniendo en cuenta que muchas de estas no cuentan con datos reales por sus restricciones legales.

La Inteligencia Artificial Generativa (IAG) en combinación con una plataforma que desarrolla con lenguajes de programación para la validación estadística, puede contribuir en este sentido, donde a través de modelos como ChatGPT pueden ofrecer alternativas para la generación de datos simulados que contemplen características de conjuntos de datos reales como parámetros estadísticos o correlaciones. Esto impacta positivamente en investigaciones y validaciones de experimentos, con datos regulados y estructurados intencionalmente.

Estos modelos son muy útiles porque pueden generar respuestas en un tiempo reducido, lo que acelera el cumplimiento de múltiples tareas y procesos. Según Feuerriegel et al. (2024), la IAG se define como técnicas computacionales capaces de generar contenido supuestamente nuevo y con gran significado. Pero estas herramientas retroalimentan sus algoritmos con datos existentes, y pueden determinar conductas y posiciones favorables para asistir a investigaciones. Lo cual ha transformado el estilo de trabajo, estudio y comunicación entre las personas. Además, Sengar et al. (2024)m refieren que al ser una herramienta que se entrena utilizando técnicas supervisadas, se ha utilizado para resolver múltiples problemas complejos en diferentes campos, demostrando tener buenos resultados. Su desarrollo evolutivo está impulsando avances significativos en diversos campos, transformando la manera en que se aborda la creación y manipulación de información en la era digital.

Esta herramienta ha tenido múltiples usos, uno de ellos es para la investigación, generando conjuntos de datos simulados. Además, se han diseñado métodos de comprobación para validar la correctitud de los datos simuladas

en correspondencia de datos reales. En López Guerrero et al. (2018), se resalta la importancia de este tipo de aplicación permite el trabajo en estudios complejos y minimiza los costos o riesgos asociados al uso de datos reales. El mismo enfoque lo expresa Vélez Torres (2019), quien describe la validación empírica como instrumento importante para garantizar la relevancia y precisión de los resultados obtenidos de un experimento. De igual forma Sánchez Vásquez et al. (2021), analizan el impacto positivo del uso de datos simulados en la formación clínica. También, Apellániz et al. (2024), proponen un enfoque para la validación de datos, desarrollando métricas para evaluar la similitud entre datos reales y simulados, y asegurando la calidad de estos. Al igual que en Marin (2024), se presenta un método para medir la similitud entre datos tabulares de pequeñas muestras y datos sintéticos generados.

Estos autores consultados, demuestran la importancia de la creación y uso de datos simulados, para la validación de experimentos y variables estadísticas. Además, la necesidad de una validación de su robustez, para lograr la confiabilidad en su consumo.

MATERIALES Y MÉTODOS

Para la simulación de datos se utilizó ChatGPT y Google Colab con Python como herramienta para la verificación de estos a través de técnicas estadísticas. Además, fue necesario implementar una metodología de trabajo, para de forma ordenada realizar el proceso de creación y comparación de datos (Figura 1).

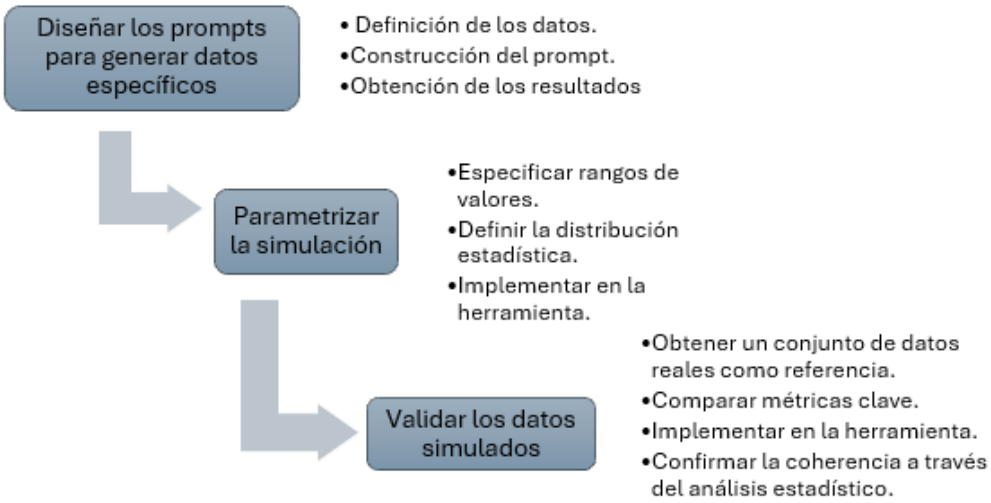

Figura 1. Metodología para el diseño de datos simulados y su validación.

Los prompts, son peticiones con estructuras entendible para la IAG, donde en el lenguaje natural del usuario, logra solicitar información, teniendo en cuenta diferentes tipos de parámetros para que se ajuste a las necesidades exigidas. Por lo que es importante diseñarlos adecuadamente. Se debe cumplir con la especificación de los tipos datos que deseas generar como número, texto, categorías. Una vez claro toda la estructura de petición a la IAG, se ejecuta en la misma para obtener los resultados. Estos resultados deben revisarse y si existe dificultades, realizar tantas iteraciones de ejecución de prompts ajustados como sea posible, hasta que se obtenga el resultado que se espera. La salida del modelo debe copiarse en un archivo de tipo csv u otro para que posteriormente sea analizado por otras plataformas.

También es muy adecuado retroalimentar parámetros en la confección de los datos simulados. Por lo que el ajuste del rango de valores, la cantidad de registro a tener en cuenta y las distribuciones estadísticas del conjunto de datos, determinan cómo se generarán estos para luego ser utilizados por otras herramientas que posibilitan su análisis y validación como por ejemplo en este trabajo se pretende implementar en Google Colab con Python.

Una vez lista la data, se procede a su validación con respecto al conjunto de datos reales para poder patentar su utilidad en la investigación. El conjunto de datos reales se obtiene de repositorios públicos o privados que los publican para la comunidad científica. Claro está, la composición de los datos reales y simulados deben ser similares para que pueda realizarse una comparación pertinente de los mismo. Se deben realizar en ambos conjuntos de datos comparaciones en cuanto a parámetros como la media, mediana, desviación estándar, percentiles y distribuciones de probabilidad. Todo esto con ayuda de Google Colab y las librerías de Python que requiera. Además, es importante corroborar coherencia a través de pruebas más complejas como T-Student o Kolmogorov-Smirnov para comparar distribuciones. Además del análisis visual no debe falta como gráficos de densidad.

Ya terminada el período de validación, se pueden realizar pruebas más complejas como regresiones, análisis

de varianza o pruebas de hipótesis. Esto permite evaluar modelos en escenarios controlados.

RESULTADOS Y DISCUSIÓN

Luego de explicar la metodología a seguir en la confección de datos simulados y todo el proceso para su validación, se realizará un caso de estudio que demuestra la aplicación de esta metodología paso a paso, para el mejor entendimiento de esta. Se utilizará Google Colab como herramienta para desarrollar los algoritmos necesarios en Python. Durante la descripción del código, se encontrarán el operador # que para Python significa un comentario de línea, que pueden incluir al reproducir el caso de estudio para documentar y que quede más explicativo.

Caso de estudio: validación de datos simulados con datos reales en el análisis de temperaturas globales

En este caso de estudio se generará los datos simulados teniendo en cuenta parámetros asociados a los datos reales utilizando herramientas como ChatGPT y Google Colab. Se verificará mediante métodos computacionales y estadísticos si estos datos simulados representan a los datos reales teniendo en cuenta la anterior metodología de trabajo descrita. Los datos reales se cargarán en Google Colab teniendo en cuenta el archivo csv "Global Land Temperatures", el mismo que se encuentra público en la plataforma Kaggle, que es muy reconocida por sus datasets validados y utilizados ampliamente en investigaciones académicas y proyectos de ciencia de datos. Estos datos pueden ser utilizados por cualquier usuario que necesite realizar experimentos con conjuntos de datos reales tratados y validados.

Pasos del Experimento:

1. Abrir Google Colab. Crear un nuevo notebook y asegurarse de que el entorno está configurado para Python 3. Subir el dataset real a la herramienta para poder trabajar con él temporalmente.
2. Debe importarse las librerías necesarias para el trabajo:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import ttest_ind
```

Comentarios de ayuda al código:

- Pandas (pd): Se utiliza para trabajar con datos estructurados en forma de tablas (DataFrames y Series).
- NumPy (np): Es una biblioteca para cálculos numéricos y manejo de arreglos multidimensionales.
- Matplotlib (plt): Se utiliza para crear visualizaciones, como gráficos de líneas, barras, histogramas y otros.
- Seaborn (sns): Es una biblioteca basada en Matplotlib que simplifica la creación de gráficos estadísticos y visualizaciones más atractivas.
- ttest_ind de SciPy: es una biblioteca para realizar análisis estadísticos avanzados.

3. Cargar los datos reales en las variables requeridas para usar en la herramienta:

```
# Cargar el dataset real desde el archivo CSV

datos_reales = pd.read_csv('GlobalTemperatures.csv')

# Filtrar datos relevantes (temperatura media anual)

datos_reales = datos_reales[['dt', 'LandAverageTemperature']]

#.loc[], que evita el warning al realizar asignaciones sobre el DataFrame

datos_reales.loc[:,'dt'] = pd.to_datetime(datos_reales['dt'])

datos_reales['Year'] = datos_reales['dt'].dt.year

datos_anuales= datos_reales.groupby('Year')['LandAverageTemperature'].mean().reset_index()
```

4. Generar los datos simulados teniendo en cuenta los parámetros de la data real

```
years = datos_anuales['Year']
```

Se calcula los parámetros poblacionales para tener en cuenta en los datos simulados desde los datos reales. Esto serían el promedio, la media y la cantidad de datos.

```
promedio = datos_anuales['LandAverageTemperature'].mean()

desviacion = datos_anuales['LandAverageTemperature'].std()

cantidad = len(datos_anuales)
```

Luego, para generar los datos simulados con una distribución normal, teniendo en cuenta el promedio, la desviación y la cantidad de datos calculado, se puede hacer con IAG. Se debe crear un prompt claro para ChatGPT. El prompt debe incluir la descripción del problema, las estadísticas necesarias: media, desviación estándar, cantidad de datos y el formato esperado de salida obtenidos de Python en el paso anterior. El prompt ejecutado para el caso de estudio fue:

"Genera una lista de datos simulados que verifique una distribución estadística normal con las siguientes características:

- Media: 8.37

- Desviación estándar: 0.58

- Tamaño del conjunto de datos: 266

Devuelve los datos en un formato tabular, con dos columnas: "Year" (1750 a 2015, en incrementos de 1 año) y "SimulatedTemperature" (los datos generados). Asegúrate de que la simulación siga la distribución normal especificada y que los datos estén en un formato adecuado para procesar en Python (CSV)."

Los datos de la media, la desviación y el tamaño total se obtiene de los análisis realizados en los datos reales. Una vez ejecutado el prompt, copia los datos simulados generados y guárdalos como un archivo CSV que se llamará "***archivo_generado_por_chatgpt.csv***" o procesa directamente en Python.

5. Para validar los datos simulados obtenidos contrastados con los reales, se realiza sus análisis descriptivos, obteniendo las características principales de estos mediante métricas clave como el promedio, desviación estándar, y percentiles.

```
# Cargar los datos simulados generados por ChatGPT

datos_simulados = pd.read_csv("archivo_generado_por_chatgpt.csv ")
```

Ya una vez cargado los datos generados por ChatGPT se realiza el análisis descriptivo pertinente:

```
# Estadísticas descriptivas

print("Datos Reales:\n", datos_anuales.describe())

print("Datos Simulados:\n", datos_simulados.describe())
```

Cuando esta sección de código se ejecutó, el resultado arrojado en la herramienta fue (Figura 2):

```
Datos Reales:
                Year  LandAverageTemperature
count   266.000000              266.000000
mean   1882.500000                8.369337
std      76.931788                0.584921
min    1750.000000                5.779833
25%    1816.250000                8.081563
50%    1882.500000                8.372167
75%    1948.750000                8.704167
max    2015.000000                9.831000
Datos Simulados:
                Year  SimulatedTemperature
count   266.000000            266.000000
mean   1882.500000              8.358843
std      76.931788              0.579136
min    1750.000000              6.473453
25%    1816.250000              7.952475
50%    1882.500000              8.386468
75%    1948.750000              8.723490
max    2015.000000             10.622880
```

Figura 2. Análisis descriptivo de los datos reales y simulados.

La interpretación de estos datos indica similitudes y diferencias entre ambas datas. Se puede constatar que el promedio de datos reales, identificado como mean, es 8.369 y el de datos simulados es 8.359. Esto significa que la similitud entre estos define que los datos simulados están alineados con la tendencia central de los datos reales.

La desviación estándar identificado como std de los datos reales es 0.5849 y de los datos simulados es 0.5791. Esto indica que la proximidad en la dispersión de los datos simulados replica bien la variabilidad observada en los datos reales.

Ambos conjuntos cubren el mismo rango de años (min:1750 a max:2015), lo que asegura consistencia en el marco temporal.

El mínimo de la temperatura de los datos reales es 5.7798 y de los datos simulados es 6.4735. El máximo de la temperatura de los datos reales es 9.8310 y de los datos simulados es 10.6229. Esto significa que los datos simulados tienen un rango un poco mayor en ambas medidas lo que pudo haber sido introducido por el modelo desarrollado. Los percentiles muestran también diferencias, aunque pequeñas, especialmente en los valores extremos (25% y 75%).

Luego se procede a utilizar herramientas de visualización de distribuciones para poder constatar si los datos simulados tienen una forma de distribución similar a los datos reales (simetría, picos, colas). También permite detectar discrepancias importantes, como distribuciones asimétricas o presencia de valores atípicos. Es una herramienta para reforzar los análisis numéricos descriptivos y estadísticos. Las representaciones gráficas como histogramas, gráficos de densidad o diagramas de caja (boxplots) hacen más accesible la interpretación de los datos a públicos no expertos.

```
sns.kdeplot(datos_anuales['LandAverageTemperature'], label='Reales', fill=True)

sns.kdeplot(datos_simulados['SimulatedTemperature'], label='Simulados', fill=True)

plt.legend()

plt.title("Comparación de Distribuciones de Temperaturas")

plt.xlabel("Temperatura Media")

plt.ylabel("Densidad")

plt.show()
```

Comentarios de ayuda al código:

kdeplot: Es una función de Seaborn que genera un gráfico de densidad.

label='Reales' y label='Simulados': Define el nombre para identificar esta distribución en la leyenda del gráfico.

fill=True: Rellena el área debajo de la curva de densidad, lo que facilita la visualización de diferencias.

legend: Agrega una leyenda al gráfico para distinguir entre las curvas de datos reales y simulados, utilizando los nombres definidos en el label.

title: Define el título del gráfico.

xlabel: Etiqueta el eje x como "Temperatura Media", indicando que mide este eje.

ylabel: Etiqueta el eje y como "Densidad".

show: Muestra el gráfico en pantalla.

Cuando esta sección de código se ejecutó, el resultado arrojado en la herramienta fue:

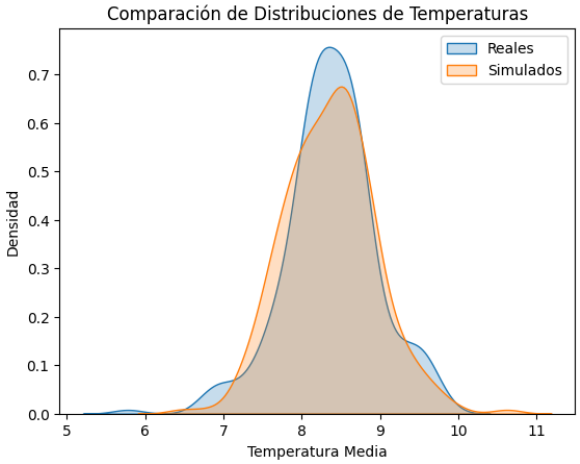

Figura 3. Comparación de distribuciones de temperatura.

La figura 3 visualiza la distribución de ambos conjuntos de datos, se puede observar que tienen una forma aproximadamente a la curva normal (campana simétrica de Gauss). Esto sugiere que las temperaturas medias reales y simuladas siguen un patrón similar en términos de variabilidad y concentración alrededor de su media. La superposición entre las curvas de los datos reales (azul) y los simulados (naranja) indica que los datos simulados reproducen bien las características generales de los datos reales, incluyendo la tendencia central y la dispersión.

Hay una semejanza en la dispersión (desviación estándar) lo cual se demuestra visualmente por la amplitud de las distribuciones. Las colas de las curvas también muestran que los valores extremos están dentro de rangos comparables, aunque los simulados presentan una ligera mayor extensión hacia valores altos (mayores de 10).

Para la validación estadística rigurosa desarrolla una prueba T-Student que permite comparar las medias de dos conjuntos de datos (en este caso, los datos reales y los simulados) para determinar si existen diferencias significativas entre ellos. Si la prueba t no encuentra diferencias significativas, se puede inferir que los datos simulados reflejan de manera adecuada las características centrales de los datos reales.

```
# Realizar prueba t

t_stat, p_value = ttest_ind(datos_anuales['LandAverageTemperature'], datos_simulados['SimulatedTemperature'])

print(f"T-statistic: {t_stat}, P-value: {p_value}")

# Interpretación del p-value

if p_value > 0.05:

 print("No hay diferencias significativas entre los datos reales y simulados.")

else:

 print("Existen diferencias significativas entre los datos reales y simulados.")
```

Al ejecutar esté código arroja como resultado:

```
T-statistic: 0.2079397847159083, P-value: 0.8353558629472315
No hay diferencias significativas entre los datos reales y simulados.
```

Figura 4. Resultado de ejecuta la prueba de t-Student.

A la hora de interpretar los resultados (Figura 4) se puede decir que el valor T-statistic igual a 0.2079 representa la diferencia en términos del error estándar entre las medias de los dos conjuntos de datos (reales y simulados). Si el valor tiende a 0, como en este caso, indica que las medias de ambos grupos son muy similares. También el valor P-value es superior a 0.05, en el caso que se analiza igual a 0.8354, no se rechaza la hipótesis (H0: Las medias de los datos reales y simulados son iguales). Esto significa que no hay diferencias estadísticamente significativas entre los datos reales y los datos simulados.

El resultado de este caso de estudio refleja que los datos simulados hallados mediante este procedimiento, constituye una representación adecuada de los datos reales y pueden utilizarse con confianza para experimentos adicionales, análisis o validaciones. Se comprobó, que la constitución de nuevos datos utilizando los parámetros poblacionales de los datos reales, los algoritmos adecuados y al mismo tiempo la IAG, generaron un conjunto de datos que emulan correctamente las propiedades estadísticas de los datos reales.

En el análisis obtenido por las herramientas utilizadas, se logró evidenciar limitaciones que pudieran constituir el punto de partida para el perfeccionamiento de estos algoritmos y procedimientos. Es lógico pensar que mientras más claras sean las instrucciones que se le dé a la IAG se obtendrán datos con mayor calidad. Aunque la IAG puede aplicar técnicas para lograr datos con distribución

normal y otras, los datos obtenidos no son idénticos a los datos que puedan ser generados por métodos estadísticos más rigurosos y por tanto con más tiempo requerido para su obtención. Por lo que puede suceder que los datos generados no representen eventos especiales o fuera del rango normal ya que como se evidencian estos se agrupan más cerca de la media, con menos valores extremos en comparación con los datos reales.

Hay que tener en cuenta la constitución de los datos reales previo a la simulación, pues pudiera caracterizarse por errores de medición, o subrepresentación de períodos, o datos fuera de rango sin sentido, lo cual se describe como sesgos con respecto a la realidad. Este tipo de características pueden ser heredadas al conjunto de datos simulados.

CONCLUSIONES

Se obtuvo un conjunto de datos simulados con patrones similares a los datos referenciados, lo que patenta que la integración de la IAG con modelos como ChatGPT y la plataforma Google Colab con Python como lenguaje de programación mostraron una alternativa que facilita la validación.

Los métodos de validación estadísticos, como T-Student, permitieron mitigar las inconsistencias generadas por las diversas corridas de los prompts en ChatGPT.

En general, la combinación de estas herramientas para el análisis representa un avance significativo en la integración de tecnologías de inteligencia artificial generativa con herramientas analíticas tradicionales. Este enfoque tiene el potencial de expandirse a diversos contextos de investigación, siempre que se implementen estrategias sólidas de validación y se consideren los sesgos inherentes a los datos generados.

REFERENCIAS BIBLIOGRÁFICAS

Apellániz, P. A., Jiménez, A., Borja Arroyo, G., Parras, J., & Zazo, S. (2024). Synthetic Tabular Data Validation: A Divergence-Based Approach. ***IEEE Access***, *12*. https://doi.org/10.1109/ACCESS.2024.3434582

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. ***Business & Information Systems Engineering***, *66*(1), 111-126. https://doi.org/10.1007/s12599-023-00834-7

Garrido Elustondo, S., Cabello Ballesteros, L., Galende Domínguez, I., Riesgo Fuertes, R., Rodríguez Barrientos, R., & Polentinos Castro, E. (2012). Investigación y protección de datos personales en atención primaria. ***Atención Primaria***, *44*(3), 172-177. https://doi.org/10.1016/j.aprim.2011.02.009

Gordon, B. G. (2020). Vulnerability in Research: Basic Ethical Concepts and General Approach to Review. ***Ochsner Journal***, *20*(1), 34-38. https://doi.org/10.31486/toj.19.0079

Kwok, C. S., Muntean, E.-A., Mallen, C. D., & Borovac, J. A. (2022). Data Collection Theory in Healthcare Research: The Minimum Dataset in Quantitative Studies. ***Clinics and Practice***, *12*(6), 832-844. https://doi.org/10.3390/clinpract12060088

López Guerrero, M. M., López Guerrero, G., & Rojano Ramos, S. (2018). Uso de un simulador para facilitar el aprendizaje de las reacciones de óxido-reducción. Estudio de caso Universidad de Málaga. ***Educación Química***, *29*(3), 79-98. https://doi.org/10.22201/fq.18708404e.2018.3.63728

López, R. G. (2024). ***Biblioguías: Datos de investigación: Los datos de investigación***. Datos de investigación: Los datos de investigación. https://uah-es.libguides.com/c.php?g=664167&p=5165387

Machuca Martínez, F. (2020). Importancia de los datos científicos y su publicación como artículo de datos. ***Ingeniería y Competitividad***, *22*(1). https://doi.org/10.25100/iyc.v22i1.8843

Marín Arraiza, P., Puerta Díaz, M., & Vidotti, S. G. (2019). Gestión de datos de investigación y bibliotecas: Preservando los nuevos bienes científicos. ***Hipertext.net***, *19, 13-31*. https://doi.org/10.31009/hipertext.net.2019.i19.02

Marin, J. (2024). ***Evaluating Synthetically Generated Data from Small Sample Sizes: An Experimental Study***, arXiv. https://doi.org/10.48550/arXiv.2211.10760

Morillo Moreno, J. C. (2024). ***Guías de la BUH: Datos de investigación: Beneficios de la gestión de datos de investigación***. https://guiasbuh.uhu.es/datosinvestigacion/beneficios

Sánchez Vásquez, U., Daniel Guerrero, A. B., Méndez Gutiérrez, E., Morales López, S., Tovar Lozano, I. I., Martínez-Rodríguez, M. A., Uribe-Campos, I. E., Sánchez-Vásquez, U., Daniel-Guerrero, A. B., Méndez-Gutiérrez, E., Morales-López, S., Tovar-Lozano, I. I., Martínez-Rodríguez, M. A., & Uribe-Campos, I. E. (2021). Diseño, elaboración y validación de un simulador realista y de bajo costo para exploración cardiaca. ***Gaceta médica de México***, *157*(1), 25-29. https://doi.org/10.24875/gmm.20005688

Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). ***Generative Artificial Intelligence: A Systematic Review and Applications***. arXiv. https://doi.org/10.48550/arXiv.2405.11029

Vélez Torres, Á. (2019). Modelación y simulación basada en agentes en ciencias sociales: Una aproximación al estado del arte. *Polis (Santiago)*, *18*(53), 282-308. https://doi.org/10.32735/s0718-6568/2019-n53-1392

Vilches, C. (2024). *Biblioguias: Gestión de datos de investigación: Protección de los datos*. Gestión de Datos de Investigación. https://biblioguias.cepal.org/c.php?g=495473&p=4398118