|  |  |
| --- | --- |
|  | **IMPUTACIÓN DE VALORES FALTANTES EN UN DATASET DE PELÍCULAS Y PROGRAMAS DE TELEVISIÓN DE NETFLIX CON PREDOMINIO DE VARIABLES CATEGÓRICAS** |
| Albino Huertas Eder Alberto1, Huillca Alarcón Edson Joel2, Niño Alcántara Ulises Emilio3, Toledo Chávez, Sally Damaris4  1. Estudiante del programa de Maestría en Ciencia de Datos en la Universidad Continental.  2. Estudiante del programa de Maestría en Ciencia de Datos en la Universidad Continental.  3. Estudiante del programa de Maestría en Ciencia de Datos en la Universidad Continental.  4. Estudiante del programa de Maestría en Ciencia de Datos en la Universidad Continental. |

**RESUMEN**

Este estudio tiene como finalidad imputar los datos faltantes en la variable *country* del conjunto de datos de Netflix, utilizando técnicas estadísticas para mejorar la calidad y consistencia de la información. Se empleó el conjunto de datos "netflix\_titles.csv", que contiene 8,807 registros sobre películas y series disponibles en la plataforma. Se realizó un análisis exploratorio para identificar la cantidad de valores nulos en la variable *country*, encontrando 831 registros incompletos, equivalentes al 9.4% del total. Para completar estos valores, se aplicó la técnica de Imputación por Grupo, reemplazando los datos faltantes con la categoría más frecuente (Moda) por cada grupo. Esta técnica fue seleccionada debido a la naturaleza categórica de la variable y la baja proporción de datos ausentes. Los resultados indicaron que la imputación permitió mantener la estructura original de la distribución geográfica del contenido sin introducir sesgos significativos. Se identificó que la mayoría del contenido proviene de países como Estados Unidos, India y Reino Unido, mientras que otras regiones tienen menor representación. La conclusión sugiere que la imputación por moda es una estrategia efectiva para completar valores faltantes en variables categóricas, asegurando la coherencia del conjunto de datos y permitiendo su uso en estudios posteriores sobre la distribución del contenido en la plataforma.

***Palabras clave***: imputación de datos, moda, análisis exploratorio, Netflix, distribución geográfica.

**ABSTRACT**

This study aims to impute the missing data in the *country* variable of the Netflix dataset using statistical techniques to improve the quality and consistency of the information. The dataset used, *netflix\_titles.csv*, contains 8,807 records of movies and series available on the platform. An exploratory analysis was conducted to identify the number of null values in the *country* variable, finding 831 incomplete records, equivalent to 9.4% of the total. To fill in these missing values, the *Group Imputation* technique was applied, replacing the missing data with the most frequent category (Mode) within each group. This technique was selected due to the categorical nature of the variable and the low proportion of missing data. The results indicated that the imputation preserved the original structure of the geographical distribution of content without introducing significant biases. It was identified that most of the content comes from countries such as the United States, India, and the United Kingdom, while other regions have lower representation. The conclusion suggests that mode imputation is an effective strategy for completing missing values in categorical variables, ensuring dataset consistency and enabling its use in further studies on content distribution on the platform.

***Keywords*:** data imputation, mode, exploratory analysis, Netflix, geographical distribution.

**INTRODUCCIÓN**

El análisis de datos frecuentemente enfrenta el problema de valores faltantes, lo que puede comprometer la calidad y validez de los resultados obtenidos (Little & Rubin, 2019). Para mitigar este problema, existen diversas técnicas de imputación, incluyendo eliminación de registros, imputación por media o mediana, y modelos más complejos como algoritmos de aprendizaje automático (Van Buuren, 2018). Sin embargo, cuando un conjunto de datos presenta un alto contenido de variables categóricas, la imputación por moda se convierte en una estrategia eficiente y adecuada, ya que permite conservar la distribución de los datos sin introducir valores atípicos o sesgos significativos (Allison, 2002).

La imputación por moda consiste en reemplazar los valores ausentes con la categoría más frecuente dentro de cada variable, lo que mantiene la coherencia de los datos categóricos y evita la pérdida de información (Enders, 2010). Además, es un método computacionalmente eficiente, lo que resulta particularmente útil en entornos con grandes volúmenes de datos donde técnicas más sofisticadas pueden ser costosas en términos de tiempo y recursos (Schafer, 1997). En comparación con otros métodos, como la imputación por regresión o árboles de decisión, la imputación por moda no requiere suposiciones complejas sobre la distribución de los datos y es menos susceptible a errores sistemáticos cuando las categorías están bien distribuidas (Van Buuren, 2018).

Dado que el dataset analizado en este estudio contiene un alto número de variables categóricas, la imputación por moda se justifica como una estrategia adecuada para el tratamiento de valores faltantes. Su simplicidad, bajo costo computacional y capacidad para preservar la estructura original de los datos refuerzan su aplicabilidad en este contexto.

En este artículo, tiene como objetivo mostrar una metodología para imputar valores faltantes de una variable categórica ‘country’ en el dataset ‘nettflix\_titles.csv’ con 12 variables a través de la moda, garantizando su integridad y calidad de datos.

**METODOLOGÍA**

**Tipo de estudio y diseño**: Es de tipo aplicativo, descriptivo y correlacional, y su diseño es no experimental y transversal con un enfoque cuantitativo.

**Muestra**: Se empleó una muestra de 8,807 registros, y se eligió imputar los valores missing de la columna “country” (con 9.44% de datos faltantes) por ser relevante para el negocio.

**Variables**: La variable objetivo es ‘country’.

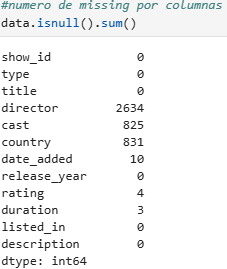
**Procedimiento de recojo de datos**: Los datos fueron conseguidos de la plataforma kaggle obtenidos del siguiente enlace:

<https://www.kaggle.com/datasets/shivamb/netflix-shows/data>

**Tratamiento y análisis de datos**:

Se identificó la necesidad de tratar los valores faltantes en el dataset de Netflix para mejorar la calidad de los datos y permitir análisis más robustos. El objetivo principal es implementar técnicas de imputación que preserven la integridad y utilidad de los datos.

Se realizó un análisis exploratorio de los datos para identificar la variables con mayor cantidad de valores faltantes, su distribución y su relación con otras variables. Se detectó que la variable categórica ‘country’ presenta valores faltantes significativos.



**Ilustración 1.** Valores faltantes por cada variable .

Se preprocesaron los datos, aplicando técnicas como codificación de variables categóricas y normalización de datos. Se diseñaron estrategias para la imputación de valores faltantes, considerando métodos como la imputación por moda, modelos basados en K-Nearest Neighbors (KNN) y algoritmos de clasificación.

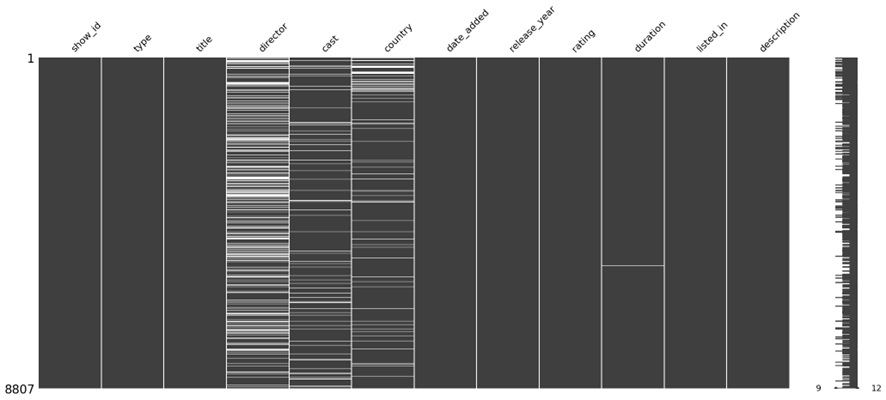
Se implementaron y compararon diferentes técnicas de imputación evaluando su eficacia en función de la precisión y la preservación de la estructura de los datos.

Los modelos de imputación fueron evaluados mediante métricas como la exactitud, la completitud y la consistencia de los datos imputados. Se seleccionó el método que mejor se adaptó a las características del dataset y a los objetivos del estudio.

Los datos imputados se integraron en el dataset final, el cual está listo para su uso.

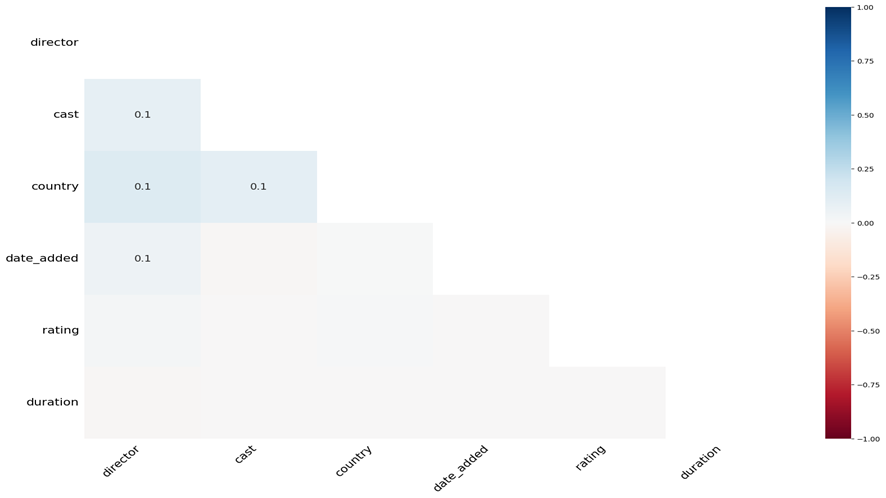
**RESULTADOS**

Se identificó que las variables con valores faltantes significativos son: ‘director’, ‘cast’ y ‘country’. Se eligió imputar los valores de la variable ‘country’.



**Ilustración 2.** Análisis de valores faltantes.

Posteriormente se realizó la matriz de correlación, de la cual se desprenden valores máximos de 0.1, lo que significa que la relación lineal entre ellas es muy débil y prácticamente inexistente, descartándose métodos de imputaciones basadas en regresiones.



**Ilustración 3.** Análisis de correlación de variables numéricas

**Se realizó la evaluación de Métodos de Imputación para Datos Categóricos**

**Métodos de Imputación:**

1. **Imputación por el Grupo**

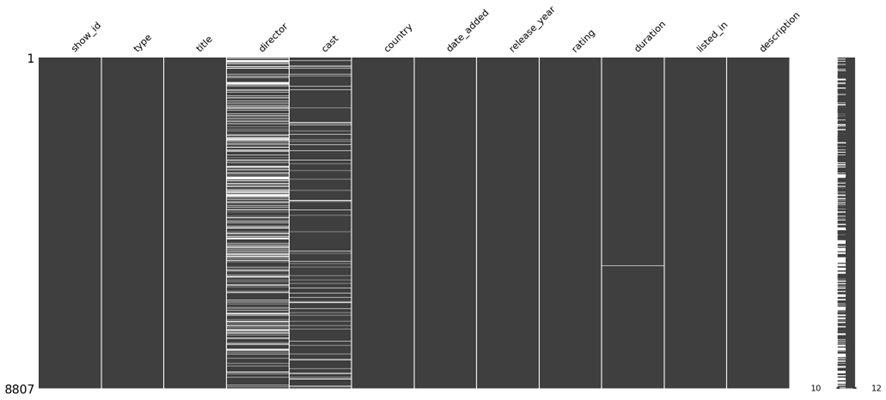
Si hay otras columnas que pueden estar referenciadas con "country" ("type" o "rating"), se pueden imputar los valores faltantes basándose en la moda de "country" dentro de esos grupos.

*Ventajas:* Proporciona una imputación más precisa al considerar el contexto de los datos.

*Desventajas*: Requiere que haya suficientes datos en los grupos para hacer una imputación válida.

*Valores imputados:*

Para este caso se ha referenciado con la columna “type”, de acuerdo al tipo de contenido: película o serie



**Ilustración 4.** Análisis de valores por imputación por grupo.

1. **Imputación por Modelo Predictivo**

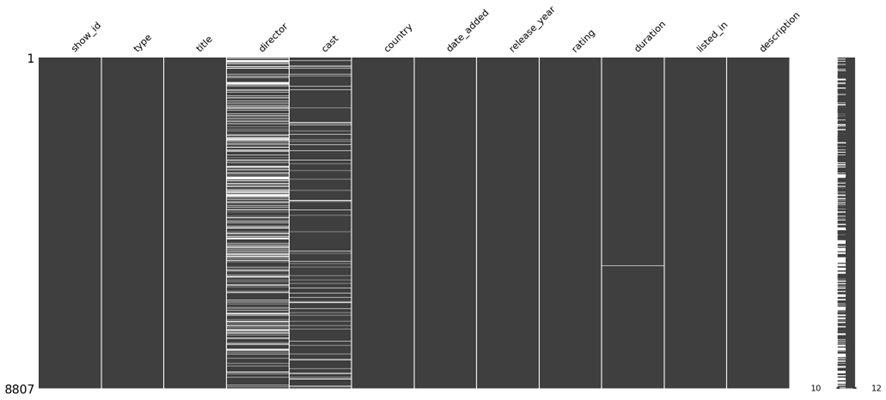
*Descripción:* El uso de algoritmos de machine learning para predecir el valor de "country".

*Ventaja:* Alta precisión si el modelo está bien entrenado.

*Desventaja:* Mayor tiempo y recursos computacionales.

*Valores imputados*: Se utilizó el modelo predictivo de KNN, sobre la misma variable bajo el concepto del vecino más cercano, para esto se transformaron los valores categóricos en numéricos, hallando el vecino más cercano, para luego volver a transformar los valores numéricos en categóricos.

*Importante:* El modelo se ha realizado sobre la misma variable, ya que no hay una correlación fuerte con otra variable para incluirla en el modelo, como ya se mencionó anteriormente

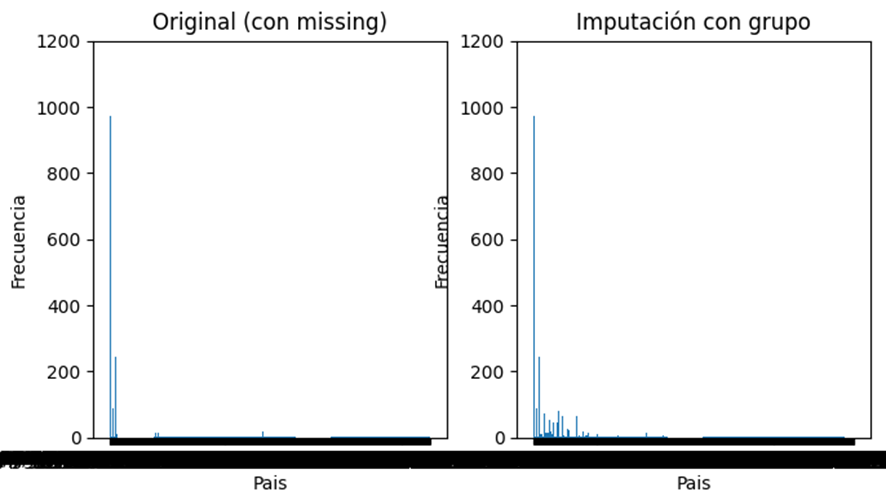


**Ilustración 5.** Análisis de valores por imputación por modelo predictivo.

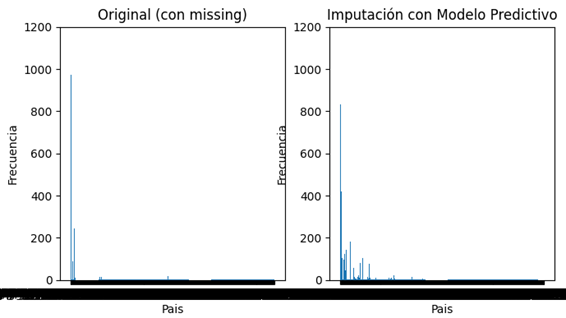
* Posteriormente se seleccionó el mejor método de imputación, de acuerdo a las siguientes validaciones:

1. **Por Distribución**

Usamos gráficos de barras para comparar la frecuencia de la columna “country” antes y después de la imputación.



**Ilustración 6.** ‘Contry’ antes vs después de la imputación con grupo.

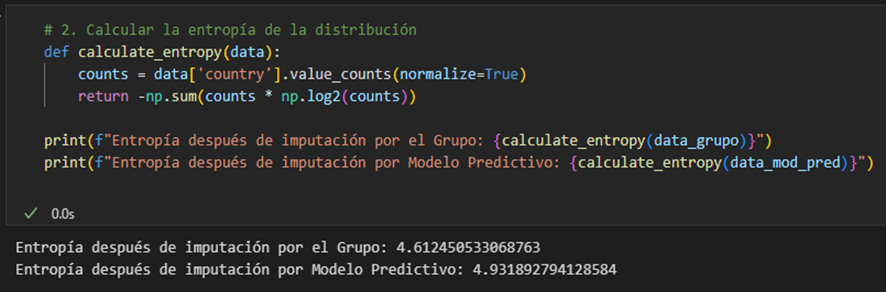


**Ilustración 6.** ‘Contry’ antes vs después de la imputación con modelo predictivo.

Se optó por preservar el método de imputación que mantenga una distribución similar a la original de los datos.

1. **Por Entropía**

La entropía mide la diversidad de los datos en la columna “country”. A mayor entropía, más desorden e incertidumbre en los datos de la columna “country”.



**Ilustración 6.** Entropía de la distribución.

*Conclusión de la verificación:*

De las verificaciones por Distribución y Entropía, se concluye que el **Método de Imputación por el Grupo,** es el **más adecuado** para imputar los datos de la columna “country”, debido a las siguientes razones:

* De acuerdo a la Validación Por Distribución, se observa que la distribución de la columna “country” con los datos sin imputar (Original) es similar con los datos ya imputados.
* De acuerdo a la Validación Por Entropía, se observa que el ratio es mucho menor frente al ratio por Imputación por Modelo Predictivo.

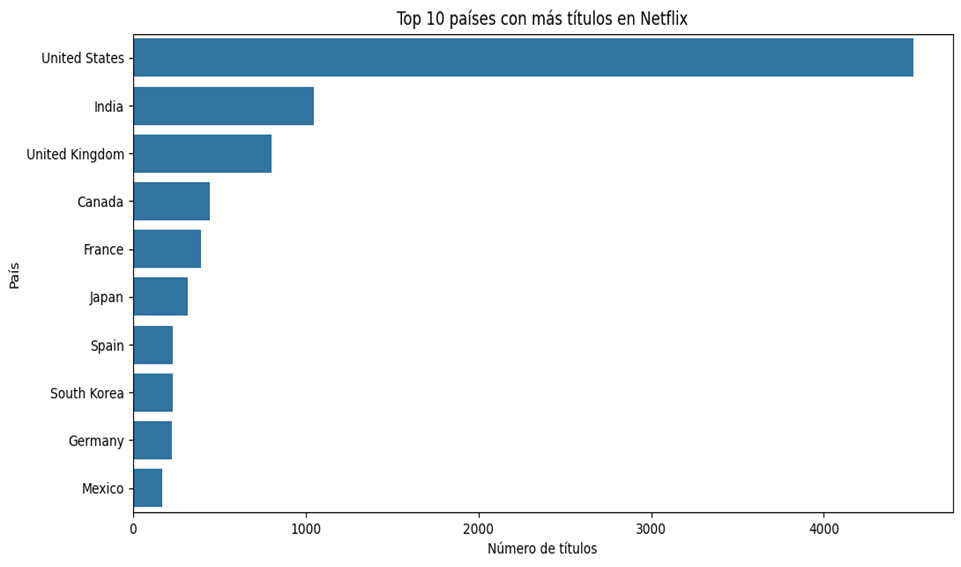
**Respondiendo las inquietudes de la Alta Dirección**

1. ***¿Qué países producen más contenido para la plataforma?***

*Cálculo del ratio:* Se determina el ratio después de haber imputado los valores para la columna ‘country’.



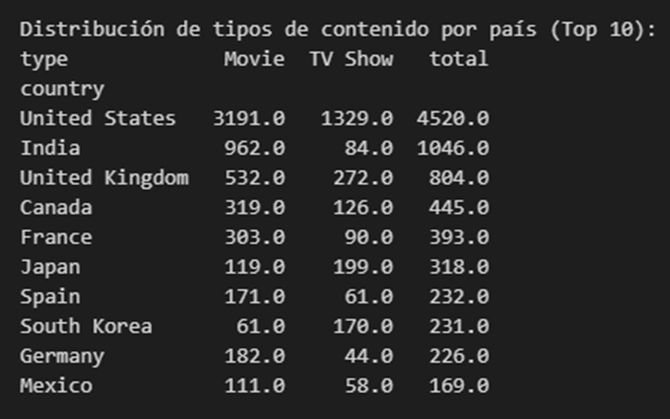
**Ilustración 7.** Países con mayor contenido en la plataforma



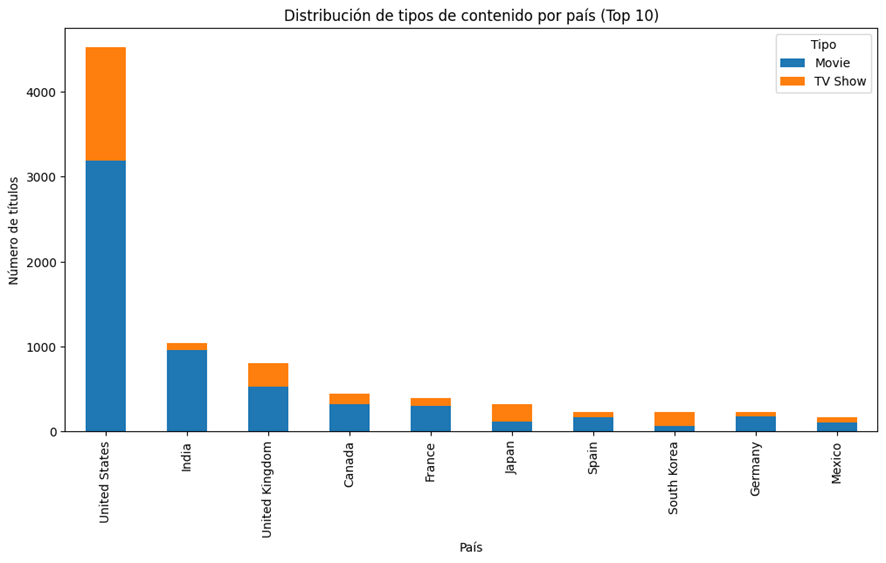
**Ilustración 8.** Gráfico de barras de países con mayor contenido en la plataforma

Según el análisis realizado, se observa que Estados Unidos es el país que más contenido produce, para la plataforma Netflix, con un total de 4,520.

1. ***¿La distribución de tipos de contenido (películas/series) por país?***



**Ilustración 9.** Países con mayor cantidad de películas y series producidas.



**Ilustración 10.** Gráfico de barras de países con mayor cantidad de películas y series producidas.

Se observa que Estados Unidos es el mayor productor de series y películas, con un total de 3,191 y 1,329 respectivamente; en segundo lugar, a nivel de series le sigue Reino Unido con 272 series producidas y en películas India con 962.

**DISCUSIÓN**

El estudio evaluó distintos métodos de imputación para datos categóricos en la variable *country* del conjunto *netflix\_titles.csv*, considerando enfoques basados en la moda dentro de grupos (*Imputación por el Grupo*) y modelos predictivos (*Imputación por KNN*). La selección del método más adecuado se realizó mediante análisis de distribución y entropía.

La *Imputación por el Grupo*, basada en la moda dentro de la columna *type* (película o serie), permitió conservar la coherencia del conjunto de datos al asignar valores según patrones existentes en Netflix. Esta técnica demostró ser eficaz al mantener una distribución similar entre los datos originales y los imputados, evitando alteraciones significativas en la representatividad geográfica del contenido.

Por otro lado, la *Imputación por KNN* utilizó el vecino más cercano tras convertir las categorías en valores numéricos. Si bien este método puede ser preciso en otros contextos, su aplicación en este caso presentó limitaciones debido a la falta de correlaciones fuertes con otras variables, lo que podría introducir mayor variabilidad en la imputación. Además, la validación por entropía mostró que este enfoque generaba mayor incertidumbre en los datos.

Los resultados indicaron que el método de *Imputación por el Grupo* era el más adecuado, ya que preservaba la estructura de los datos originales sin aumentar el desorden. Esto es crucial, considerando que los países con más títulos (Estados Unidos, India y Reino Unido) son mercados clave para Netflix. Además, los países con menos títulos podrían representar oportunidades estratégicas para la expansión del catálogo y la atracción de nuevos usuarios.

Este análisis proporciona una visión detallada de la distribución geográfica del contenido en Netflix, lo que permite a la alta dirección tomar decisiones informadas sobre estrategias de localización y diversificación del catálogo, optimizando así su oferta de contenido a nivel global.

**COMPARACIÓN CON OTRO CASO DE ESTUDIO**

Se realiza la comparación con el siguiente caso de estudio: ***“Imputación de Datos en Series de Precipitación Diaria - Caso de Estudio Cuenca del Río Quindío”***, que se encuentra en la siguiente dirección web:

<https://dialnet.unirioja.es/descarga/articulo/5478781.pdf>

1. **Similitudes:**
2. *Objetivo de la imputación:* Ambos estudios buscan mejorar la calidad de los datos al imputar valores faltantes en variables clave (país en el caso de Netflix y precipitación en el caso hidrológico).
3. *Comparación de métodos:* Ambos análisis comparan múltiples técnicas de imputación y seleccionan la más adecuada según ciertos criterios.
4. *Impacto de la imputación en la interpretación de datos:* En ambos casos, se destaca la importancia de preservar la estructura y distribución de los datos originales.
5. *Relevancia de la validación:* Se emplean métricas específicas para validar los resultados: en el estudio de Netflix, se usa la distribución y la entropía, mientras que en el de precipitación, se comparan estadísticos de las series originales e imputadas.
6. **Diferencias:**
7. *Tipo de datos:* El análisis de Netflix trabaja con datos categóricos (países), mientras que el estudio hidrológico maneja datos numéricos continuos (precipitación diaria).
8. *Contexto:* La imputación en Netflix busca mantener patrones de distribución geográfica de contenido, mientras que en el caso hidrológico, se busca preservar la coherencia climática.
9. *Métodos utilizados:* En Netflix, los métodos incluyen imputación por grupos y KNN, en la precipitación, se usan interpolaciones ponderadas, con énfasis en el método de la Medida Estadística Ponderada.
10. Limitaciones y desafíos específicos: En Netflix, el problema principal con KNN fue la falta de correlaciones fuertes entre variables, en precipitación, la imputación fue efectiva con datos faltantes menores al 20 %, pero problemático con datos faltantes superiores (ejemplo: estación Barragán con 48,21 %).

* ***Se concluye que***, ambos estudios abordan el problema de la imputación de datos desde perspectivas diferentes (Datos Categóricos vs. Numéricos) y en contextos distintos (Entretenimiento vs. Hidrología). Sin embargo, comparten un enfoque metodológico similar en la evaluación de múltiples métodos de imputación y la importancia de preservar la integridad de los datos originales. Las diferencias radican en la naturaleza de los datos, los métodos utilizados y las aplicaciones prácticas de los resultados.

**Contribuciones de los autores:** Todos los autores del presente artículo aportaron con la codificación en Python, la búsqueda de bibliografía, así como la redacción y revisión.

**Fuentes de financiamiento:** autofinanciado.

**Conflictos de interés:** los autores declaran no tener conflictos de interés en la publicación del artículo.

**Agradecimientos:** A la dedicación de los mismos autores y muy especial al docente por las indicaciones y sugerencias.

**BIBLIOGRAFÍA**

**Little, R. J., & Rubin, D. B. (2019).** *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

**Allison, P. D. (2002).** *Missing Data* (Vol. 136). Sage Publications.

**Acock, A. C. (2005).** "Working with Missing Values." *Journal of Marriage and Family, 67(4), 1012-1028.*

**James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).** *An Introduction to Statistical Learning* (2nd ed.). Springer.

**Kang, H. (2013).** "The prevention and handling of the missing data." *Korean Journal of Anesthesiology, 64(5), 402-406.*

**Sharma, M., & Kumar, V. (2021).** "A Review on Imputation Techniques for Handling Missing Data." *Materials Today: Proceedings, 45(5), 4852-4857.*

**Newman, M. E. J. (2018).** *Networks: An Introduction* (2nd ed.). Oxford University Press.