

DEPÓSITO LEGAL ZU2020000153

ISSN 0041-8811

E-ISSN 2665-0428

Revista de la Universidad del Zulia

Fundada en 1947
por el Dr. Jesús Enrique Lossada



Ciencias
Exactas,
Naturales
y de la Salud

77
ANIVERSARIO

Año 15 N° 43
Mayo - Agosto 2024
Tercera Época
Maracaibo-Venezuela

Métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas en perspectiva comparada

Edwin Fernando Mejía-Peñañiel*
Johanna Enith Aguilar-Reyes**
Paulina Fernanda Bolaños-Logroño***
Jorge Rigoberto López-Ortega****

RESUMEN

La imputación por regresión, la imputación por moda, la imputación múltiple y los árboles de decisión son métodos utilizados para tratar valores faltantes en variables categóricas. En este contexto, el objetivo de la investigación consiste en definir los criterios para comparar estos métodos, estableciendo en el proceso sus semejanzas y diferencias conceptuales. En lo metodológico, se hizo uso del método comparativo y del análisis epistemológico como condición de posibilidad para entender el alcance y significado de cada uno de estos métodos, con arreglo al acopio de fuentes documentales de comprobado valor científico. Aunque parciales y limitados, los resultados obtenidos respaldan la hipótesis de que la estadística es una ciencia en desarrollo que requiere del diálogo inter-metodológico, como condición de posibilidad para obtener resultados más precisos, incluso cuando faltan ciertos datos en una ecuación. En última instancia, este trabajo sienta las bases para futuras investigaciones que podrían profundizar en la perspectiva comparada de ciertos métodos y herramientas de investigación, como los que se enuncian en el título del artículo.

PALABRAS CLAVE: Estadística, imputación por regresión, imputación por moda, imputaciones múltiples, árboles de decisión.

*Docente en la Escuela Superior Politécnica de Chimborazo, Riobamba – Ecuador. ORCID: <https://orcid.org/0000-0001-6888-4621>. E-mail: efmejia@epoch.edu.ec

** Docente en la Escuela Superior Politécnica de Chimborazo, Riobamba – Ecuador. ORCID: <https://orcid.org/0000-0002-1230-2503>. E-mail: johannae.aguilar@epoch.edu.ec.

***Docente en la Escuela Superior Politécnica de Chimborazo, Riobamba – Ecuador. ORCID: <https://orcid.org/0000-0003-3911-0461>. E-mail: paulina.bolanos@epoch.edu.ec

****Docente Investigador en el Grupo Ciencia de Datos CITED - Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba - Ecuador. ORCID: <https://orcid.org/0000-0002-4790-6876>. E-mail: jorge.lopezo@epoch.edu.ec

Recibido: 20/02/2024

Aceptado: 12/04/2024

Methods of Imputation by Regression, Imputation by Mode, Multiple Imputation and Decision Trees for Categorical Variables in Comparative Perspective

ABSTRACT

Regression imputation, mode imputation, multiple imputation and decision trees are methods used to deal with missing values in categorical variables. In this context, the objective of the research is to define the criteria for comparing these methods, establishing in the process their conceptual similarities and differences. Methodologically, use was made of the comparative method and epistemological analysis as a condition of possibility to understand the scope and significance of each of these methods, based on the collection of documentary sources of proven scientific value. Although partial and limited, the results obtained support the hypothesis that statistics is a developing science that requires inter-methodological dialogue as a condition of possibility to obtain more accurate results, even when certain data are missing in an equation. Ultimately, this work lays the groundwork for future research that could deepen the comparative perspective of certain research methods and tools, such as those stated in the title of the article.

KEYWORDS: Statistics, regression imputation, mode imputation, multiple imputations, decision trees.

Introducción

Normalmente los métodos por imputación se utilizan para datos faltantes en análisis estadísticos, con énfasis en el caso de variables categóricas. En este contexto, los métodos de imputación se dividen en un conjunto de sub-métodos con procedimientos particulares, tales como: a) Imputación múltiple: el cual es apropiado en la estadística bayesiana y la información de la muestra para realizar inferencia respecto de los parámetros. Tal como sostienen Alfaro y Fuenzalida (2009), es ventajoso cuando la variable a imputar tiene características particulares como en el caso de las variables discretas.

Del mismo modo, destaca dentro del menú de las opciones propias de los métodos por imputación la llamada imputación por regresión, método que en su decurso utiliza modelos de regresión para imputar los valores faltantes. Es adecuado para variables cuantitativas y permite

relaciones no lineales entre las variables explicativas y la variable dependiente (Medina & Galván, 2007).

Por su parte, destaca también la imputación por moda que precisamente imputa los valores faltantes con la frecuencia más común de la variable y; los árboles de decisión, que son esencialmente algoritmos de aprendizaje automático que se utilizan para clasificación de datos, incluidos los datos con variables categóricas. Tendencialmente, son eficientes para categorizar variables numéricas y permiten además relaciones no lineales entre las variables explicativas y la variable dependiente. En palabras de Minguillon y Pujol:

Entre los sistemas de clasificación los árboles de decisión, los árboles de decisión constituyen uno de los métodos más utilizados, por su simplicidad y por su facilidad de construcción. No obstante, usualmente han sido sustituidos por otros métodos (redes neuronales, *support vector machines*...) debido en parte a su dependencia de los conjuntos de datos utilizados en el entrenamiento y debido también a su limitada capacidad de producir resultados (2015, p. 28).

De cualquier manera, es difícil determinar *a priori* el valor analítico de una herramienta como los árboles de decisión, todo dependerá en último término de la naturaleza de la operación estadística o algorítmica que se quiera realizar y, más aún, de los datos disponibles. Además de estas consideraciones en la práctica, la elección del método de imputación depende del tipo de variable a imputar y del contexto del análisis. Finalmente, es importante tener en cuenta que las imputaciones pueden introducir cierta incertidumbre en los resultados y que ciertamente la imputación de datos faltantes no siempre es la mejor opción.

En este contexto epistemológico, es decir, en el marco de la filosofía de la ciencia que reflexiona sobre el alcance y significado del conocimiento científico en general y sobre la viabilidad de los métodos, teorías y técnicas de una disciplina en particular (Bunge, 2005), el objetivo de la investigación consiste en definir los criterios para comparar estos métodos, estableciendo en el proceso sus semejanzas y diferencias conceptuales más destacadas.

El artículo se divide en 5 secciones, en la primera se describen las bases teóricas de la investigación en términos de la literatura revisada selectivamente. Seguidamente, se explican los fundamentos metodológicos del artículo. En la tercera sección, por su parte, se exponen las características generales del tema abordado, para luego analizar los principales resultados

obtenidos, sin ninguna pretensión erudita de arribar a verdades generales. Finalmente, se presentan como un insumo válido para el debate y la discusión estadística, las principales conclusiones de la investigación, en el entendido que se trata de una indagación parcial y limitada en su alcance.

1. Balance teórico de la investigación

El presente apartado tiene un doble propósito, por un lado, aclara más allá de toda duda razonable el sentido y significado que tienen para los autores de la investigación, los principales conceptos y categorías de análisis que componen al tema, mucho más cuanto que, la estadística es una disciplina de síntesis epistemológica que no está exenta de debates en su fundamentos metódicos y conceptuales, de hecho, ninguna ciencia lo está. Por el otro, se muestran al lector informado las principales influencias teóricas que hicieron posible el desarrollo de este artículo.

En principio y tal como sostienen (Radío, 2017; Alfaro & Fuenzalida, 2009) la imputación es el proceso de estimar y reemplazar los valores faltantes en un conjunto de datos. Existen diversos métodos de imputación, cada uno con sus propias ventajas y desventajas. A continuación, se describen de forma somera los métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión, en el sentido que son asumidos por los autores de este artículo y en la forma como son descritos por la literatura científica de mayor divulgación.

En líneas generales, la imputación por regresión, también conocida como método Buck, utiliza un modelo de regresión para estimar los valores faltantes en una variable dependiente basándose en las variables explicativas disponibles. Tal como explican Useche y Mesa (2006) este método fue propuesto por primera vez por Buck en el año 1960 y se ha utilizado ampliamente en el campo de la estadística y la investigación con diversos resultados.

Como complemento de la regresión, la imputación por moda es un método o herramienta simple que reemplaza los valores faltantes en una variable categórica con el valor de la categoría más frecuente en esa variable, lo que, llegado el caso, puede servir para mostrar un panorama más o menos completo de las variables que se quieren medir, como condición de posibilidad para cuantificar de forma racional determinados fenómenos (Bello, Cuta, & García,

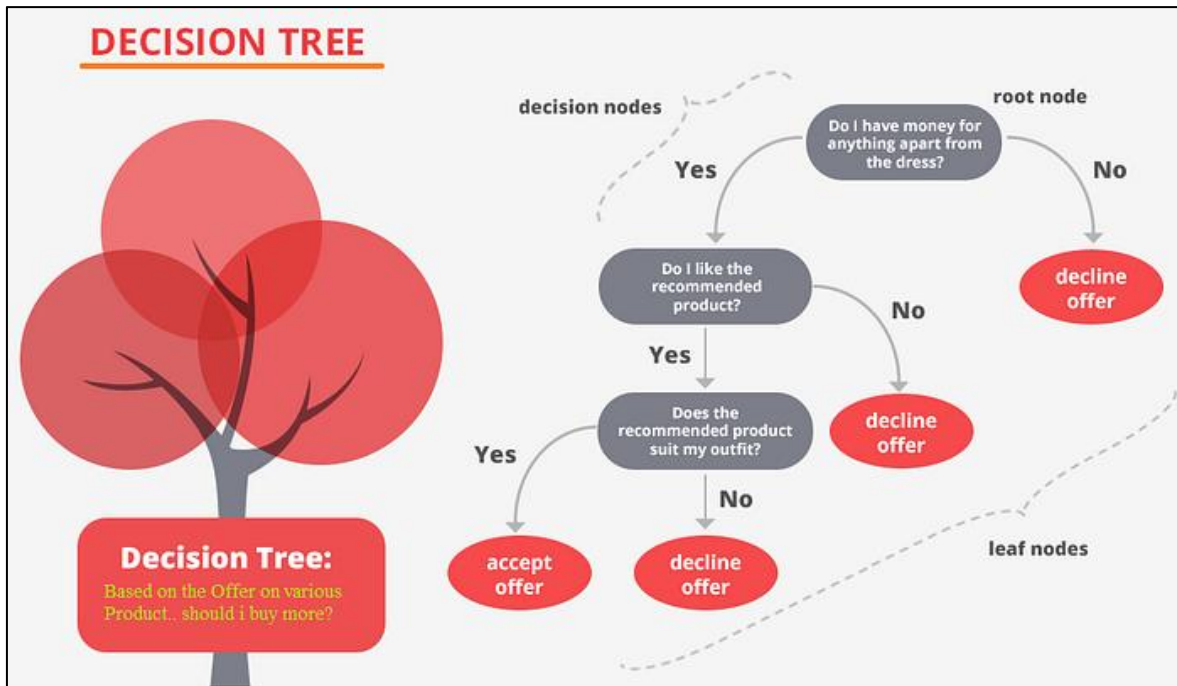
2019) . Aunque es un método sencillo, puede ser adecuado en algunas situaciones, especialmente cuando los datos faltantes están distribuidos de manera aleatoria y no hay una relación clara entre las variables (Rivas, Martínez, & Galindo, 2010) ; no obstante, conviene preguntar entonces ¿Hasta qué punto se pueden sustituir por complete ciertos datos?

Por su parte, la imputación múltiple es un método que crea varias versiones completas de un conjunto de datos con valores faltantes, cada una con diferentes estimaciones de los valores faltantes. Luego, se combinan los resultados de estas versiones para obtener una estimación final de los parámetros de interés estadístico. Definitivamente, la imputación múltiple puede ser más precisa que otros métodos de imputación, ya que tiene en cuenta la incertidumbre asociada con la estimación de los valores faltantes, tal como sostienen Muñoz y Alvarez (2009).

En cuanto a los árboles de decisión son una técnica de aprendizaje automático que se puede utilizar para la imputación de valores faltantes en variables categóricas o continuas. En este método, se construye un árbol de decisión que divide el conjunto de datos en subconjuntos homogéneos en función de las variables explicativas disponibles. Luego, se imputan los valores faltantes en cada subconjunto utilizando diferentes métodos, como la imputación por la media o la mediana (Minguillon & Pujol, 2015). Empero, como ya se dijo en la introducción se trata de una técnica esquemática que, como los mapas semánticos o el análisis FODA dependen para su correcta realización de la capacidad analítica y creativa del autor del estudio (Rivas, Martínez, & Galindo, 2010).

Para los efectos particulares de este marco teórico, también fue importante la revisión de la literatura sobre la Imputación basada en árboles de clasificación, como técnica que utiliza árboles de decisión, entre otros dispositivos analíticos, para imputar valores faltantes en variables categóricas. En este método, se divide la población respondiente en dos subconjuntos aleatoriamente, y se utiliza el primer subconjunto para construir el árbol de decisión y, el segundo subconjunto, para aplicar la imputación, Normalmente, esta técnica puede reducir el error de imputación al evitar el error que se puede cometer al imputar mediante información obtenida de los registros que han participado en la construcción del árbol (Alfaro & Fuenzalida, 2009; Medina & Galván, 2007; Muñoz & Alvarez, 2009)

Cuadro No. 01: Modelo conceptual de árbol de decisión



Fuente: Tomado de Medium (2020).

De la interpretación de las fuentes recabadas se desprenden al menos tres conclusiones básicas que conviene analizar: primero, existen diversos métodos de imputación que pueden ser utilizados para tratar los valores faltantes en un conjunto de datos. Segundo, la elección del método adecuado depende del tipo de variable, la distribución de los datos faltantes y el objetivo del análisis. Tercero, la imputación por regresión, la imputación por moda, la imputación múltiple y los árboles de decisión son algunos de los métodos más comunes que se utilizan en la práctica de los estadísticos, sin embargo, no son las únicas opciones ni están libres de errores o contradicciones intersecas a su propia naturaleza conceptual.

2. Metodología

Como se dijo desde el resumen, se hizo uso del método comparativo y del análisis epistemológico como condición de posibilidad para entender el alcance y significado de la imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión, en particular. Entendemos el método comparativo como una herramienta básica propia de las ciencias sociales para develar, en este caso particular, las semejanzas y diferencias de diferentes

herramientas metodológicas en el marco de la estadística. Nuestra visión de la comparación metódica es subsidiaria de los aportes de Piovani & Krawczyk (2017), quienes se preguntan:

¿A qué nos referimos cuando hablamos de un acto de comparación? En el lenguaje ordinario, siguiendo una clásica definición de diccionario, se presenta la comparación como el acto de ‘observar dos o más cosas para descubrir sus relaciones o estimar sus diferencias y semejanzas’. En el lenguaje epistemológico, se define como una operación intelectual a través de la cual se cotejan los estados de uno (o más) objetos sobre la base de al menos una propiedad común. (2017, p. 823)

Sin lugar a dudas, interesaron a los autores de este artículo la dimensión epistemológica de la comparación, en la cual se cotejan los estados de objetos, ideales o reales, que tienen algunas propiedades en común de conformidad con su naturaleza diferencial. En este orden de ideas, el análisis epistemológico es una herramienta típica de la filosofía de la ciencia que busca comprender la esencia y existencia de saberes, conceptos, teorías o incluso metodologías, en el contexto de la ciencia, disciplina o filosofía donde se producen y reproducen, bajo la impronta de determinados intereses, científicos o incluso meta-científicos.

Tal como observan algunos epistemólogos consagrados, es decir, académicos que estudian la teoría del conocimiento científico, entre los que destacan: Bunge (2005), Dancy (1993) y Llano (1991) y otros. La producción científica, incluida la estadística, no puede sustraerse del tiempo y espacio donde sucede como un fenómeno intelectual susceptible a la investigación racional, que puede revelar al entendimiento, no solo su alcance y significado epistémico, sino además, las condiciones políticas, económicas y sociales en las cuales estos dispositivos se producen, nunca como elementos neutrales, sino como saberes condicionados por prácticas, creencias, intereses e ideologías, razón por la cual el análisis epistemológico si bien se origina en la filosofía es, por derecho propio, una forma de análisis histórico-holístico.

Aclarado lo anterior, conviene enfatizar que lo que aquí se hace no abarca toda la dimensión histórica de lo que debe significar un verdadero análisis epistemológico, a lo sumo, los autores de la investigación se conformaron con comparar las semejanzas y diferencias esenciales de los métodos señalados, sin llegar a agotar toda su complejidad gnoseológica, trabajo que por su dificultad intrínseca sobrepasa los límites del formato artículo científico.

3. Métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas en perspectiva comparada

En este apartado, que representa metafóricamente el plato fuerte de la investigación, se presenta el análisis comparativo de los métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas, de forma ordenada, con especial énfasis en los aspectos centrales de cada método.

3.1. Imputación por regresión

La imputación por regresión utiliza un modelo de regresión para predecir los valores faltantes a partir de los valores observados de otras variables. Esencialmente, es una técnica sencilla y rápida de implementar, pero puede introducir sesgos si los supuestos del modelo de regresión no se cumplen. Por estas razones, funciona mejor para variables numéricas que para variables categóricas.

3.2. Imputación por moda

La imputación por moda reemplaza los valores faltantes con el valor más frecuente (la moda) de esa variable. Es un método simple y rápido, pero puede distorsionar la distribución original de los datos. Normalmente, funciona mejor para variables categóricas que para variables numéricas.

3.3. Imputación múltiple

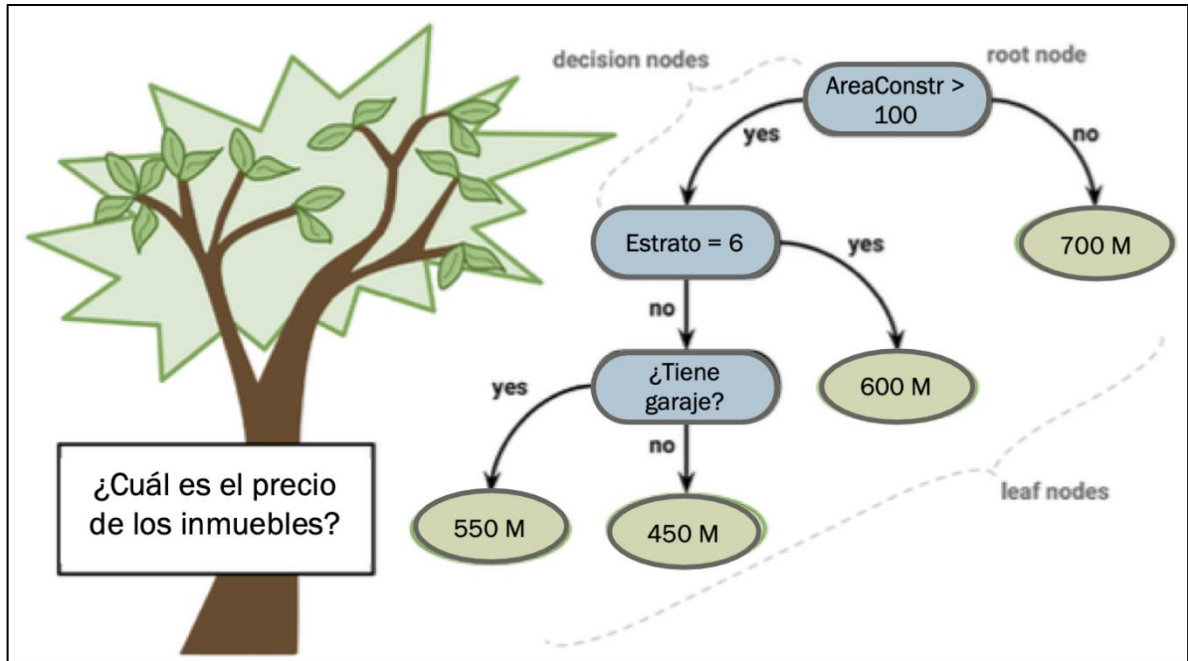
La imputación múltiple genera múltiples conjuntos de datos imputados, ajusta un modelo a cada uno, y combina los resultados para obtener estimaciones finales. Es más compleja de implementar que los métodos anteriores, pero produce estimaciones menos sesgadas, especialmente cuando los datos no son *Missing Completely At Random* (MCAR). Definitivamente, puede aplicarse tanto a variables numéricas como categóricas.

3.4. Árboles de decisión

Los árboles de decisión son modelos de aprendizaje automático que pueden manejar tanto variables numéricas como categóricas sin necesidad de transformaciones. Pueden capturar interacciones complejas entre predictores sin requerir supuestos sobre la distribución

de los datos. Por estas razones, son más favorables a valores atípicos que otros métodos. Pueden lidiar con datos faltantes al realizar "particiones sustitutas" cuando una variable predictora tiene un valor faltante.

Cuadro No. 02: Adaptación de la estructura de un árbol de decisión al campo de la valoración inmobiliaria



Fuente: Tomado de (Guijarro, 2023).

En resumen, la imputación múltiple y los árboles de decisión son los métodos más flexibles y robustos para manejar variables categóricas con datos faltantes, mientras que la imputación por regresión y por moda tienen limitaciones cuando se trabaja con este tipo de variables. A continuación, se presenta un cuadro comparativo que pretende ilustrar las semejanzas y diferencias de cada método.

4. Discusión de resultados

En perspectiva comparada, los métodos de imputación por regresión, imputación por moda e imputación múltiple son, más allá de sus diferencias y particularidades, métodos utilizados para completar valores perdidos en datos, mientras que los árboles de decisión son utilizados para clasificar y predecir resultados basados en datos existentes (Bello, Cuta, &

García, 2019; Medina & Galván, 2007). Por su parte, la imputación por regresión es un método paramétrico que utiliza modelos de regresión para predecir valores faltantes en variables numéricas o categóricas, asumiendo una relación lineal entre las variables independientes y la variable a imputar.

Cuadro No. 01: Comparativa general de los métodos objeto de estudio

Método	Función central del método	Semejanzas	Diferencias	Observaciones
Imputación por regresión	La imputación por regresión utiliza un modelo de regresión para predecir los valores faltantes a partir de los valores observados de otras variables.	La imputación por regresión tiende a ser más precisa al utilizar modelos para predecir valores faltantes, mientras que la imputación por moda y, los demás métodos abordados, son más simples, pero la regresión puede no reflejar la realidad de los datos.	En líneas generales, la imputación por regresión, puede mantener mejor la distribución original de los datos, mientras que los otros métodos, más allá de sus particularidades, pueden distorsionarla al imputar valores comunes.	La imputación por regresión es más compleja al requerir la construcción de modelos, mientras que la imputación por moda es más directa y fácil de implementar.
Imputación por moda	La imputación por moda implica reemplazar los valores faltantes con el valor más común en una variable. Como se sabe, es un método simple y rápido de imputación, pero al mismo tiempo adecuado para variables categóricas o discretas.	Al igual que los demás métodos, la imputación múltiple implica generar múltiples conjuntos de datos completos, cada uno con valores imputados de manera diferente, para reflejar la incertidumbre en los datos faltantes.	La imputación múltiple tiende a ser más precisa al considerar la incertidumbre y generar múltiples estimaciones, mientras que la imputación por moda es más simple, pero puede ser menos precisa al imputar un único valor común.	En contraste con los otros métodos, la imputación múltiple es más compleja al generar múltiples conjuntos de datos completos, mientras que, por ejemplo, la imputación por moda es más directa y fácil de implementar. En definitiva, la imputación múltiple es más sofisticada y puede ofrecer estimaciones más robustas al considerar la incertidumbre en los datos faltantes, mientras que la imputación por moda es más simple,

				pero puede ser menos precisa al imputar un único valor común. La elección entre ambos métodos dependerá de la naturaleza de los datos y la importancia de reflejar la incertidumbre en las estimaciones.
Imputación múltiple y árboles de decisión	El método de imputaciones múltiples se utiliza para completar valores perdidos en los datos, especialmente en variables categóricas. Esta técnica utiliza múltiples conjuntos de datos completos para generar una única estimación, lo que mejora, al menos en teoría, la precisión de los resultados y permite incorporar información de otras variables (Imputación múltiple en variables categóricas usando data augmentation y árboles de clasificación, 2010)	A semejanza de los árboles de decisión, la imputación múltiple se utiliza para comprender procesos complejos de toma de decisiones y mejorar la toma de decisiones en diferentes ámbitos de la vida en sociedad. Los árboles de decisión son especialmente útiles cuando la variable objetivo es categórica, es decir, pertenece a un conjunto discreto de clases o categorías (Novoa-Hernández, Cobos-Valdes, Samaniego-Mena, & Novoa-Pérez, 2018).	En sentido estricto, la diferencia de los árboles de decisión en comparación con los otros métodos abordados, no es en ningún caso categórica o antagónica. Se trata más bien de una herramienta que puede complementar analíticamente a otros métodos. Por ejemplo, en el contexto de variables categóricas, el método de imputaciones múltiples puede ser utilizado para completar valores perdidos antes de aplicar los árboles de decisión. La imputación múltiple permite manejar datos incompletos y generar estimaciones más precisas, lo que puede mejorar la calidad de los datos y la eficiencia del análisis de los árboles de decisión	En definitiva, el método de imputaciones múltiples y los árboles de decisión son dos enfoques útiles en el análisis de datos, especialmente en el contexto de variables categóricas. La imputación múltiple puede ser utilizada para completar valores perdidos y mejorar la calidad de los datos, mientras que los árboles de decisión pueden ser utilizados para comprender procesos complejos de toma de decisiones y mejorar la toma de decisiones racionales, es decir, aquellas que son basan en el balance de la relación costo-beneficio (Minguillon & Pujol, 2015).

Fuente: Elaboración propia con arreglo al objetivo general de la investigación.

La imputación por moda es un método simple que completa valores perdidos con la frecuencia más común de la variable. En contraste, la imputación múltiple es un método que combina varios conjuntos de datos completos para generar estimaciones más precisas de valores faltantes. No obstante, en el caso de variables categóricas, los árboles de decisión son útiles para clasificar y predecir resultados basados en datos existentes. Definitivamente, los árboles de decisión utilizan algoritmos de aprendizaje automático para dividir los datos en función de su capacidad para separar la variable objetivo de manera efectiva (Novoa-Hernández, Cobos-Valdes, Samaniego-Mena, & Novoa-Pérez, 2018)

Tal como demostró el cuadro comparativo, los métodos de imputación por regresión, imputación por moda e imputación múltiple son útiles para completar valores perdidos en datos, mientras que los árboles de decisión son útiles para clasificar y predecir resultados basados en datos existentes. Ambos enfoques tienen sus propias ventajas y desafíos y se pueden utilizar en diferentes escenarios según las necesidades del análisis de datos y las capacidades del equipo de investigación.

Conclusiones

Los resultados de nuestro análisis epistemológico en perspectiva comparada sostienen que, los métodos de imputación, como la imputación múltiple, la imputación por moda y los árboles de decisión, son importantes, o en determinados casos fundamentales, en el análisis de datos para completar valores perdidos y mejorar la calidad general de los datos. Más específicamente la imputación múltiple proporciona un análisis de los patrones de datos perdidos y se utiliza para generar imputaciones múltiples, lo que permite el análisis de conjuntos de datos completos con procedimientos que admiten conjuntos de datos de imputación múltiple y vinculaciones inter metodológicas.

Por su parte, la imputación por moda es una técnica simple que subsana en lo posible los valores perdidos con la frecuencia más común de la variable. En este marco general, los árboles de decisión son útiles para comprender procesos complejos de toma de decisiones y mejorar la toma de decisiones, ya que permiten dividir los datos en función de su capacidad para separar la variable objetivo de manera efectiva. En conclusión, estos métodos son importantes para

completar datos perdidos y mejorar la calidad de los datos, lo que resulta útil en el análisis de datos y la toma de decisiones estadísticamente informada.

Por lo demás, los resultados obtenidos en este estudio, aunque parciales y limitados, respaldan la hipótesis de que la estadística es una ciencia en desarrollo que requiere del dialogo inter-metodológico, como condición de posibilidad para obtener resultados más precisos, incluso cuando faltan ciertos datos en una ecuación.

Se ha demostrado de manera concluyente que herramientas como el análisis epistemológico y el método comparativo tienen mucho que aportar para campos como la estadística. Estos hallazgos no solo confirman la importancia de una visión renovada de las ciencias exactas, sino que también abren nuevas vías de investigación y posibles aplicaciones prácticas en las que se construyen puentes epistemológicos entre las ciencias sociales y las ciencias exactas. En última instancia, este trabajo sienta las bases para futuras investigaciones que podrían profundizar en la perspectiva comparada de ciertos métodos y herramientas de investigación como la: imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas ampliar nuestro entendimiento de estos temas y otros similares.

Finalmente, los resultados presentados aquí tienen el potencial de impactar positivamente en el entendimiento epistemológico de los saberes y métodos que estructuran a la estadística, mucho más relevante cuanto que: “Los métodos estadísticos son ampliamente utilizados en diferentes áreas del quehacer humano: el gobierno, los negocios, la educación, la psicología, la sociología, la antropología, las ciencias del comportamiento, la agricultura, la medicina, la biología y la física, entre otras” (Porrás, 2019, p. 02).

Referencias

Alfaro, R., & Fuenzalida, M. (2009). Imputación múltiple en encuestas microeconómicas. *Nota Técnica*, (46), 273-288. <https://www.scielo.cl/pdf/cecon/v46n134/art07.pdf>.

Bacallao Guerra, Jorge & Bacallao Gallestey, Jorge (2010). Imputación múltiple en variables categóricas usando data augmentation y árboles de clasificación. *Revista Investigacion Operacional*, 31 (02), 133-139. En: <file:///D:/Descargas/Dialnet-ImputacionMultipleEnVariablesCategoricasUsandoData-3218968.pdf>

Bello, A., Cuta, J., & García, E. (2019). Técnicas de imputación para datos de precipitación máxima mensual en la zona central de Boyacá. *Revista Ingeniería, Investigación y Desarrollo*, 19 (01), 64-79.

Bunge, M. (2005). *La ciencia. Su método y su filosofía*. Buenos Aires: Debolsillo.

Dancy, J. (1993). *Introducción a la epistemología contemporánea*. Madrid: Tecnos.

Guijarro, F. (12 de marzo de 2023). *Rpub*. Obtenido de Árboles de decisión: <https://rpubs.com/fraguima/lonja5>

Llano, A. (1991). *Gnoseología*. Barañáin-Pamplona: Ediciones de la universidad de Navarra.

Medina, F., & Galván, M. (2007). *Imputación de datos: Teoría y práctica*. Santiago de Chile. CEPAL. Serie 54. En: <https://repositorio.cepal.org/server/api/core/bitstreams/02dd479f-fae2-43c4-b5ec-5419fa7f6190/content>

Medium (01 de septiembre de 2020). *Árboles de decisión (Práctica)*. Obtenido de Bootcamp AI: <https://bootcampai.medium.com/árboles-de-decisión-práctica-62ee5c578b08>

Minguillon, J., & Pujol, J. (2015). *Arboles de decisión. Terceras Jornadas de Matemática Discreta y Algorítmica*. En: <https://idus.us.es/bitstream/handle/11441/75448/Árboles%20de%20decisión.pdf?sequence=1&isAllowed=y>

Muñoz, J., & Alvarez, E. (2009). Métodos de imputación para el tratamiento de datos faltantes: Aplicación mediante R/S plus. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, (07), 03-30.

Novoa-Hernández, P., Cobos-Valdes, D., Samaniego-Mena, E., & Novoa-Pérez, M. (2018). Árboles de decisión para la evaluación del riesgo biológico de procesos biofarmacéuticos. *Revista Ciencia Unem*, 11 (28), 08-17. <https://www.redalyc.org/journal/5826/582661251001/582661251001.pdf>.

Piovani, J., & Krawczyk, N. (2017). Los Estudios Comparativos: algunas notas históricas, epistemológicas y metodológicas. *Educação & Realidade, Porto Alegre*, 42 (03), 821-840. <http://dx.doi.org/10.1590/2175-623667609>.

Porras, A. (2019). *Diplomado en análisis de información geoespacial. Conceptos básicos de estadística*. México DF. <https://centrogeo.repositorioinstitucional.mx/jspui/bitstream/1012/157/1/13-Conceptos%20Básicos%20de%20Estadística%20%20Diplomado%20en%20Análisis%20de%20Información%20Geoespacial.pdf>: Centro publico de investigacion CONACYT.

Radío, G. R. (2017). *Los valores perdidos en el muestreo de poblaciones*. Vigo. http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1468.pdf: Univresidad de la Corunay Univresidad de Vigo.

Rivas, C., Martínez, M. d., & Galindo, P. (2010). La imputación múltiple como alternativa al análisis de la no respuesta en la variable intención de voto. *Revista Española de Ciencia Política*. Núm. 22, Marzo 2010, pp. 99-118. En: https://www.researchgate.net/publication/235979983_La_imputacion_multiple_como_alternativa_al_analisis_de_la_no_respuesta_en_la_variable_intencion_de_voto

Useche, L. M., & Mesa, D. M. (2006). Una introducción a la Imputación de Valores Perdidos. *Terra Nueva Etapa*, XXII (31), 127-151. <https://www.redalyc.org/pdf/721/72103106.pdf>.