UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

“IMPUTACIÓN DE DATOS FALTANTES EN LOS INGRESOS POR HOGAR EN LA ENAHO UTILIZANDO EL MÉTODO DEL K-VECINO MÁS CERCANO”

Presentada por:

Oscar Ronald Collazos Tuesta

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima – Perú

2021

UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"IMPUTACIÓN DE DATOS FALTANTES EN LOS INGRESOS POR HOGAR EN LA ENAHO UTILIZANDO EL MÉTODO DEL K-VECINO MÁS CERCANO"

Presentada por:

Oscar Ronald Collazos Tuesta

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

Mg.Sc. Clodomiro Fernando Miranda Villagómez **PRESIDENTE**	Dr. César Higinio Menacho Chiok **PATROCINADOR**
Mg. Jaime Carlos Porras Cerrón **MIEMBRO**	Mg. Carlos López de Castilla Vásquez **MIEMBRO**

Lima – Perú

2021

Dedicatoria

La presente Tesis está dedicada a toda mi familia, principalmente a mi mama Hortencia que ha sido fundamental en mi formación como persona y profesional, y a mi papa Mesias por brindarme su apoyo y los recursos necesarios para perseguir esta meta.

A mi esposa Mariam por sus palabras de aliento y por brindarme su amor, paciencia y comprensión, y a mis hijas Sofía y Fabiana que son mi motivación más grande.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

RESUMEN

La Encuesta Nacional de Hogares (ENAHO), es el instrumento que utiliza el Instituto Nacional de Estadística e Informática (INEI) para recopilar a nivel nacional los datos de los hogares sobre su condiciones económicas, educativas, salud, etc. y que permiten generar indicadores que miden el estado y la evolución de la pobreza, el bienestar y las condiciones de vida de los hogares del Perú, así como para efectuar diagnósticos y medir el alcance de los programas sociales (alimentarios y no alimentarios) en la mejora de las condiciones de vida de la población peruana. Sin embargo, un problema que debe enfrentar la ENAHO es la no respuesta total o parcial en las unidades de muestreo (no respuesta en unidades) o en una pregunta específica (no respuesta por ítem); sobre todo a las preguntas referidas a los ingresos de los hogares.

Para el tratamiento de los datos faltantes, se han propuesto una variedad de métodos que comprenden desde el más simple que consiste en la eliminación de las observaciones que tengan algún dato faltante en una de las variables hasta métodos más consistentes basados en un proceso de imputación con los datos faltantes a partir de los datos completos. El objetivo de esta investigación es presentar y aplicar los métodos de imputación de la media y mediana, el método Hot-Deck y el k vecino más cercano para estimar los datos faltantes del Ingreso por hogar en la ENAHO 2017 trimestre 3. Los resultados indican que los datos faltantes del ingreso tienen un mecanismo MCAR. La estimación del intervalo de confianza del 95% para la media de los ingresos imputados, tuvieron amplitudes por el método de la media 131,41 (el menor) mientras que por el k vecino más cercano fue 139,4. Para estimación de la desviación estándar del ingreso, fue el menor para la media 92,97 y k vecino más cercano 100,99. Los resultados de la comparación de los métodos de imputación, fueron usando los datos completos para generar una muestra aleatoria de datos faltantes artificiales y luego se hallaron el Cuadrado Medio del Error (ECM) y correlaciones con los datos observados e imputados para cada método. El método del k vecino más cercano tuvo los menores valores de ECM 1412,6 y 444,4 para la media y mediana; mientras que los otros métodos sus valores fueron por la media 1504,5; por la mediana 1619,9 y por el Hot-Deck 1963,7. Los coeficientes de correlaciones resultaron con valores muy similares, para k vecino más cercano 0,968 con la media y 0,964 con la mediana.

Palabras claves. Mecanismo MCAR, Imputación en la ENAHO, Imputación Hot_deck, Imputación k vecino más cercano, Package R "VIM".

ABSTRACT

The National Household Survey (ENAHO) is the instrument used by the National Institute of Statistics and Informatics (INEI) to collect national data on household economic, educational and health conditions, etc. and that allow generating indicators that measure the status and evolution of poverty, well-being and living conditions of Peruvian households, as well as to carry out diagnoses and measure the scope of social programs (food and non-food) in the improvement of the living conditions of the Peruvian population. However, a problem that ENAHO must face is the total or partial non-response in the sampling units (non-response in units) or in a specific question (non-response per item); especially to the questions referring to the income of the households.

For the treatment of missing data, a variety of methods have been proposed , ranging from the simplest, which consists of elimination of observations that have some missing data in one of the variables, to most consistent methods based on an imputation process with the missing data from the complete data. The objective of this research is to present and apply the imputation methods of the mean and median, the Hot-Deck method and the nearest k neighbor to estimate the missing data of the Income per household in the ENAHO 2017 quarter 3. The results indicate that missing income data has a MCAR mechanism. The estimate of the 95% confidence interval for the mean of the imputed income, had amplitudes by the method of the mean 131.41 (the smallest) while for the nearest k neighbor it was 139.4. To estimate the standard deviation of income, it was the lowest for the mean 92.97 and k nearest neighbor 100.99. The results of the comparison of the imputation methods, were using the complete data to generate a random sample of artificial missing data, and then the Mean Square Error (ECM) and correlations with the observed and imputed data for each method were found. The closest neighbor k method had the lowest ECM values of 1412.6 and 444.4 for the mean and median; while the other methods their values were by the average 1504.5; by the median 1619.9 and by the Hot-Deck 1963.7. The correlation coefficients resulted in very similar values, for k nearest neighbor 0.968 with the mean and 0.964 with the median.

Keywords. MCAR Mechanism, Imputation in the ENAHO, Imputation Hot_deck, Imputation k nearest neighbor, Package R "VIM".

I. INTRODUCCIÓN

En las últimas décadas muchos países están realizando importantes esfuerzos destinados a consolidar, fortalecer y ampliar los sistemas de información provenientes de sus diferentes sectores productivos, y en particular de los datos proveniente de las encuestas de hogares. En el contexto social es sumamente importante contar con información actualizada y relevante de los hogares y sus miembros, puesto que esto permite desarrollar programas sociales que permitan enfrentar los desafíos planteados para reducir la pobreza y la desigualdad, mejorar el nivel de vida y en general el bienestar de la población de un país. Sin embargo, un problema común que se presenta cuando se aplican encuestas a hogares, es la falta de respuesta total o parcial en las unidades de muestreo (no respuesta en unidades) o bien en una pregunta específica (no respuesta por ítem). Según (Cochran, 1977), la falta de respuesta en las encuestas de hogares está asociada a diversas causas como la fatiga del informante, el desconocimiento de la información solicitada, el rechazo de las personas a informar acerca de temas sensibles, la negativa de los hogares a participar en la investigación, así como a problemas asociados a la calidad del marco de muestreo.

La literatura muestra que los problemas de los datos faltantes en un conjunto de datos pueden ser considerados desde varios enfoques. Según (Polo, C., Behar, R., & Olaya, J., 2000), los principales problemas que surgen cuando se tienen datos faltantes son: a) pérdida de la eficiencia de los estimadores, b) presencia de sesgos y c) complicación en el manejo y análisis de los datos. La pérdida de la eficiencia, sucede cuando el análisis se realiza con datos sólo de las variables completas, lo cual disminuye el tamaño de la muestra teniendo un cambio en la variancia afectando la eficiencia de la precisión de las estimaciones. Los sesgos, aparecen cuando se supone que las unidades con datos faltantes son similares a los completos (es una muestra representativa), ignorando la existencia de los mismos por lo cual se produce un sesgo en las estimaciones. La complicación en el manejo y análisis ocurre al tener bases de datos incompletas que deben ser manejadas con métodos estadísticos que consideran bases de datos completas, por lo cual se debe decidir qué hacer con los datos faltantes. Para Rubin (1976), el efecto en la consistencia de los resultados por parte de los diferentes usuarios; que implica que estos pueden usar diferentes métodos para el tratamiento de los datos faltantes obteniendo resultados diferentes, perdiéndose la consistencia, uniformidad y confiabilidad de la información. Para Schafer (1997), cuando se maneja conjuntos de datos multivariados, la omisión de registros incompletos puede ser ineficiente debido a las grandes cantidades de datos que se descarta referente al conjunto de las variables que están completas.

Un aspecto que se debe considerar antes de elegir algún método para el tratamiento de datos faltantes, es determinar la distribución o el patrón de comportamiento de dichos datos; puesto que la aplicación eficiente de algunos métodos supone que la distribución de datos no faltantes posee un patrón determinado. En (Little & Rubin, D.B., 2002), menciona que los patrones de los datos faltantes se pueden clasificar según su grado de aleatoriedad en tres grupos: MCAR (valores perdidos o faltantes de manera completamente aleatoria), MAR (valores perdidos o faltantes de manera aleatoria) y NMAR (valores perdidos o faltantes que no siguen un proceso aleatorio). En el caso de MCAR la ausencia de datos no está asociada con ninguna variable presente en el conjunto de datos (faltantes y completos), mientras que en MAR la ausencia de datos está asociada a variables con datos completos y en un NMAR los datos faltantes dependen sólo de las variables con datos faltantes.

Para el tratamiento de datos faltantes existen una variedad de técnicas y métodos. Según (Sande, 1982), se puede considerar: eliminar todos los registros que tengan al menos un dato faltante, ignorar los datos faltantes en cada caso o imputar los datos faltantes. Cuando no se pueden ignorar los datos faltantes, una manera de tratarlos es reemplazarlos con valores internos o externos, a este procedimiento se le denomina imputación. En (Lohr, 1999), refiriéndose a la imputación señala que su importancia no solo radica en reducir el sesgo por la ausencia de respuestas, sino también en producir un conjunto de datos completos y consistentes que puedan ser analizados con las técnicas estadísticas. En (Batista, G. & Monard, M. C., 2002), se describen algunos de los métodos de imputación que son ampliamente utilizados. El método de imputación más simple y práctico consiste en sustituir los datos faltantes de cada variable por la media o mediana (variable cuantitativa) o la moda (variable cualitativa) para lo cual se usa sólo los valores conocidos de las variables, suponiendo un patrón MCAR. Sin embargo, tienen una serie de implicancias y desventajas; tales como: cambio de la distribución de la variable imputada, produce correlaciones con otras variables, los errores estándares son muy pequeños y estimaciones sesgadas (Schafer & Graham, J., 2002) . Los métodos Hot-Deck son una alternativa para solucionar los problemas de la imputación de la media. Dentro de los métodos Hot-Deck que proporciona una serie de ventajas para la imputación de los datos, es el uso de la técnica del k -vecino más cercano para estimar y sustituir los datos faltantes basándose en la cercanía medida a través de una métrica de distancia. Las principales ventajas son: i) el k-vecino más cercano puede imputar variables cuantitativas y cualitativas, ii) no necesita definir un modelo para los datos faltantes y iii) puede ser aplicado fácilmente para el tratamiento de valores faltantes múltiples (Batista, G. E. & Monard, M. C., 2002).

En la presente investigación se aplica el método k-vecino más cercano para imputar los valores faltantes de la variable Ingreso mensual por hogar en la ENAHO 2017 del tercer trimestre a nivel nacional. El proceso de imputación propuesto permitirá obtener una base de datos completa respecto al ingreso de los hogares.

Los objetivos de la investigación son los siguientes:

1. Presentar el proceso de imputación de datos faltantes aplicando el método del k-vecino más cercano para los ingresos de los hogares de la ENAHO.
2. Identificar el patrón de distribución de los datos faltantes del ingreso por hogar en la ENAHO 2017.
3. Obtener estimaciones por intervalo de confianza para la media y desviación estándar de los ingresos mensuales de los hogares para cada uno de los métodos de imputación.
4. Comparar el método de imputación del k vecino más cercano con los métodos de la media, la mediana y Hot-deck aleatorio mediante el Error Cuadrado Medio.

II. REVISIÓN DE LITERATURA

2.1 Manejo de los datos faltantes en las encuestas

Es muy común encontrar cuando se aplican encuestas que todas las unidades de estudio o análisis no respondan a todas las preguntas, con lo cual se origina el problema conocido como no respuesta, datos faltantes o datos missing. Las causas de estas no respuestas, pueden ser consideradas desde una amplia variedad de factores, que pueden ir desde una pregunta mal elaborada hasta una unidad de muestreo no existente. La proporción de no respuesta en las preguntas (variables) de las unidades de muestreo (observaciones) pueden variar y puede depender del estudio (la dificultad de llenar el cuestionario o de la medición de una unidad) o del propio proceso de la investigación; lo que tipifica a la unidad como una pérdida parcial o total dependiendo del número de datos faltantes. Un aspecto que se debe considerar en un conjunto de datos es determinar hasta qué porcentaje de pérdida de datos (ítems) se considera tratable mediante los métodos de imputación o considerarla como una mala recolección de datos, y por tanto, la base de datos obtenida es muy defectuosa que simplemente no debe ser usada. La respuesta a esto no es tan sencilla. En la práctica se habla de pérdidas máximas entre 1 y 25% de la data dependiendo de la exactitud del estudio y del área de investigación entre otros factores (Goicochea, 2002).

2.1.1 Fuentes y tipos de datos faltantes

Las investigaciones estadísticas pueden ser; investigaciones por muestreo, investigaciones exhaustivas, experimentos comparativos y estudios observacionales. Todas estas investigaciones son estructuradas realizando un conjunto de fases en la cual pueden estar presentes fuentes de errores que pueden originar la no respuesta (Goicochea, 2002).

- **La fase de planteamiento de la investigación.** El no entender de manera adecuada el objetivo de la investigación, lleva a una mala definición de la población, y por tanto, al uso de un instructivo de recolección de datos incorrecto, definición errónea de conceptos que traería como consecuencia que las preguntas o mediciones no sean las adecuadas con la unidad de análisis, lo que podría llevar a obtener errores en las investigaciones que dan origen a la no respuesta.

- **En la fase de elaboración de instrumentos básicos.** Se pueden presentar problemas como, preguntas del cuestionario que no reflejen los objetivos de la investigación, posiciones inadecuadas de las preguntas, mala redacción, influencia por parte del

encuestador, cuestionarios muy largos, preguntas que requieren de memoria a mediano o largo plazo o que sean comprometedoras (personales), lo que llevaría a que el encuestado no responda todas las preguntas.

- **En la fase de diseño de la encuesta.** Puede ocurrir la no respuesta, por un diseño no acorde con el objetivo, marcos imperfectos, selección de la encuesta a juicio propio, aun cuando se propone una selección probabilística.

- **En la fase de organización y ejecución de operaciones de campo.** La no respuesta depende mucho del comportamiento del encuestador, de una mala selección, mal entrenamiento, sentimientos personales del entrevistador que pueden ser transmitidos a los encuestados produciéndose errores, supervisión inadecuada, cobertura incompleta de la zona geográfica en estudio por razones de costo, tiempo muy corto o por errores de selección de las unidades de muestreo en campo.

- **Para la fase de procesamiento de datos.** puede haber ausencia o inconsistencias, debido posiblemente a un perfil o capacitación inadecuada de los codificadores / transcriptores, produciéndose errores tales como; incorrecta trascripción y codificación de los datos, utilización no adecuada de códigos y validaciones, sistemas de control de calidad deficientes y uso de “software” y “hardware” inadecuados.

- **En fase de análisis de resultados.** Resaltan más las consecuencias de la no respuesta, debido a que ésta puede conducir a errores en la misma, ya que no contiene todas las variables necesarias en el análisis, conduciendo a un cálculo inadecuado de pesos, lo que produce una distorsión de los resultados reales.

- **Plan de difusión.** Los resultados obtenidos pueden ser erróneamente publicados, presentando excesiva información que puede confundir al lector.

Las fuentes de errores presentes en las diferentes fases de una investigación estadística, pueden producir también no respuesta. Los errores que se producen pueden ser de dos tipos: errores muestrales y errores no muestrales (Goicochea, 2002).

- **Los errores muestrales.** Son los errores producidos al observar una muestra de la población y no la totalidad de ella. Este tipo de error está compuesto por la variabilidad del estimador ante muestras repetidas y su sesgo, llamado sesgo técnico (Mesa & Useche, 2006)

- **Los errores no muestrales.** Son aquellos errores presentes en una investigación, no atribuibles al observar una muestra. Pueden ser aleatorios o sistemáticos. Los aleatorios: exceso o duplicación (ocurre cuando algunas unidades de observación aparecen más de una vez en el marco de muestreo), cuando los datos que se obtiene de la unidad de observación es incorrecta. Sistemáticos: errores de cobertura (problemas en el marco muestral), por la no inclusión de algunas unidades de observación, las observaciones seleccionadas para la encuesta no proporcionan todos los datos que deberían recogerse.

Tipos de datos faltantes

Cuando se aplican encuestas a hogares, la ocurrencia de los datos faltantes o no respuesta se pueden distinguir, entre la falta de respuesta total (no se encontró al informante, se rechazó la entrevista, problemas de marco de muestreo, etc.) y la no respuesta parcial, que se asocia con situaciones en que no se obtuvo respuesta en algunas preguntas del cuestionario. La falta de respuesta total se corrige generalmente eliminando las observaciones y ajustando los factores de expansión, de modo que las unidades que permanezcan en la muestra puedan estimar sin sesgos los parámetros de la población. Por su parte, para corregir la omisión parcial es común que se aplique algún método de imputación de datos. La aplicación de los distintos métodos de imputación de datos faltantes propuestos, requieren que se conozca el patrón o mecanismo de los datos faltantes.

La no respuesta, aunque cuando no se quiera, siempre estará latente su presencia en toda investigación que involucre la medición de datos sobre individuos u hogares, y que muchas veces no se obtendrán registros completos cuyas causas diversas son ajenas al investigador. La situación idónea en una investigación es obtener una base de datos completa, con los valores reales que permita aplicar las tradicionales técnicas de análisis de datos. La no respuesta, puede presentarse de dos maneras:

- **La no respuesta total**. Es cuando falta todo el registro de una base de datos, por ausencia de la unidad a medir o por impedimento de efectuar un conjunto total de mediciones de variables en un determinado momento específico, es decir, no se recoge ningún dato de la unidad de la muestra. Por ejemplo, cuando se lleva a cabo la aplicación de encuesta en hogares y en algunas viviendas seleccionadas, y no se encuentran personas al momento de aplicar el instrumento, generándose una pérdida total de las respuestas del cuestionario que se le iba a aplicar a ese hogar o a esa persona.

- **La no respuesta parcial.** Se presenta cuando hay ausencia de una o más variables, sin llegar a la ausencia completa de un registro, ejemplo; un individuo a encuestar se encuentra, pero no responde algunas preguntas del cuestionario o a una unidad no se le efectuaron algunas mediciones, por fallas en los equipos, sin embargo, otras mediciones si se llevaron a cabo. La no respuesta parcial puede tener dos formas de presentarse; cuando las variables de un registro están ausentes, porque la data no está disponible, o cuando una variable produce una inconsistencia con el resto de las variables, entonces se dicen que están "pérdidas artificialmente" debido a que han sido eliminadas mediante un proceso de depuración, éstas últimas serán tratadas como la primera forma.

2.1.2 Mecanismos y patrones de los datos faltantes

En los estudios de encuestas por muestreo, se distingue en la no respuesta por unidad, la cual ocurre cuando la totalidad de la unidad de muestreo no registra datos (Ejemplo. Un hogar no desea contestar la encuesta); mientras que la no respuesta por ítem, ocurre cuando no se registra datos para alguna o varias variables en diferentes unidades de muestreo (Ejemplo. Una persona no responde su ingreso). El manejo estadístico para la no respuesta por unidad, se realiza por un proceso de reponderación, mientras para la no respuesta por ítem se realiza por el análisis de datos completos, datos disponibles y por métodos de imputación. La decisión del procedimiento o método más adecuado para resolver el problema de no respuesta en el ítem, implica conocer el mecanismo y patrón de la ocurrencia de los datos faltantes.

Los mecanismos de ocurrencia de los datos faltantes son asunciones acerca de la naturaleza y tipo de datos faltantes. Estos mecanismos de ocurrencia, permiten establecer la relación que puede existir entre los datos faltantes y las variables definidas. La tipología propuesta por (Rubin, 1976), es ampliamente utilizada dónde los clasifica en tres grupos según su grado de aleatoriedad y la relación entre los datos faltantes y los datos completos: Datos faltantes completamente aleatorio (MCAR: Missing Completely At Random), datos faltantes aleatorios (MAR: Missing At Random) y datos faltantes no aleatorios (NMAR: Missing Not At Random). Así mismo, se agrupan los datos faltantes por el patrón de pérdida en: ignorables y no ignorables. Los patrones de pérdida son ignorables, si los datos faltantes ocurren de manera MCAR o MAR; lo que quiere decir es que en el análisis de los datos se puede ignorar las razones de la ocurrencia de los datos faltantes. Si los datos faltantes ocurren con el patrón NMAR, se dice que son no ignorables.

En (Schafer & Schenker, Inference with Imputed Conditional Means, 1997), se define la matriz de datos $Z=(Z_{Obs}, Z_{Mis})$, dónde Z_{Obs} contienen los datos observados o completos y Z_{Mis} contienen los datos faltantes o missing. Se define la matriz indicadora (binaria) R, dónde cada elemento $R_{ij}=1$, si Z_{ij} es un dato faltante y $R_{ij}=0$, si Z_{ij} es un observado. A partir, de estas definiciones de matrices, se consigue un modelo para la distribución de probabilidad para R con la finalidad de definir los mecanismos de los datos faltantes. Entonces la matriz R está constituida por un conjunto de variables aleatorias y que depende del mecanismo de los datos faltantes MAR, MCAR y MNAR.

- **Datos faltantes completamente al azar (MCAR)**

El mecanismo de los datos faltantes es MCAR, cuando su ocurrencia no está relacionada o no dependen de ninguna de las variables del conjunto de datos (no dependen de Z_{Obs} o Z_{Mis}). Los datos faltantes podrían considerarse como un MCAR, si la distribución de R no depende de los datos observados o los datos faltantes. El MCAR significa que la probabilidad de la ocurrencia de los valores faltantes (R=1) no depende de Z_{Obs} o de Z_{Mis}. Entonces se tiene:

$$P(R=1/Z_{Obs}, Z_{Mis}) = P(R=1)$$

El mecanismo de datos faltantes MCAR, es considerado el mejor contexto para el manejo de valores faltantes sobre todo para aplicar los métodos de imputación de datos, puesto que son totalmente arbitrarios y no se producirá sesgo en las estimaciones (Little, T.D., Jorgensen, T., Lang, K., & Moore, E., 2014).

- **Datos faltantes al azar (MAR)**

Los datos faltantes corresponden a un MAR, cuando su ocurrencia sólo está relacionada o depende de las variables con datos completos (Z_{Obs}) y no de las que poseen datos faltantes (Z_{Mis}). Es decir, los datos faltantes podrían considerarse como un MAR, si la distribución de R sólo depende de los datos observados. El mecanismo MAR significa que la probabilidad de la ocurrencia de los valores faltantes (R=1) depende de Z_{Obs} y no de Z_{Mis}. Entonces se tiene:

$$P(R=1/Z_{Obs}, Z_{Mis}) = P(R=1/Z_{Obs})$$

- **Datos faltantes no al azar (MNAR)**

Los datos faltantes siguen un MNAR, cuando su ocurrencia sólo depende de las variables con datos faltantes (Z_{Mis}) y no de las que poseen datos completos (Z_{Obs}). Los datos faltantes son

considerados como un MNAR, si la distribución de R sólo depende de los datos faltantes. El MNAR significa que la probabilidad de la ocurrencia de los valores faltantes (R=1), sólo depende de Z_{Mis}. Entonces se tiene:

$$P(R=1/Z_{Obs}, Z_{Mis}) = P(R=1/Z_{Mis})$$

Ejemplo.

En una encuesta sobre la calidad de vida de los hogares, se tiene datos sobre el ingreso y el nivel socioeconómico del hogar. Suponiendo que existen datos faltantes en el ingreso del hogar. Los datos se presentan en la Cuadro 1.

Cuadro 1. Ingresos y nivel socioeconómico de los hogares (Ejemplo)

Nivel socioeconómico (X1)	Ingreso del hogar (decenas soles) (X2)			
	Datos completos	MCAR	MAR	MNAR
A	35	35	35	
A	45		45	
A	65	65	65	65
A	85	85	85	85
B	50			
B	70	60		70
B	90			90
B	45	45		
C	55	55	55	
C	80	80	80	80
C	55		70	
C	95	95	95	95

Los datos faltantes en la variable ingreso, seguirá un mecanismo MCAR cuando la ocurrencia de dichos datos faltantes no depende (es independiente) del nivel socioeconómico y del ingreso. En la columna MCAR, los datos faltantes del ingreso aparecen en cualquier de los niveles socioeconómicos (A, B o C) y para cualquier valor del ingreso (bajos o altos), indicando que los datos faltantes no dependen (no están relacionados) con el nivel socioeconómico y el ingreso. En este caso, los datos faltantes son considerados un MCAR. La probabilidad que ocurra un dato faltante en el ingreso, no dependerá del nivel socioeconómico ni del mismo ingreso. Esto es:

$$P\ (Ingreso = Missing)/Nivel, Ingresos) = P\ (Ingreso = Missing)$$

Para el caso que los datos faltantes en la variable ingreso tenga un mecanismo MAR, la ocurrencia de dichos datos faltantes sólo dependerá del nivel socioeconómico y no del ingreso. En la columna MAR, los datos faltantes del ingreso aparecen sólo para el nivel socioeconómico B y para cualquier valor del ingreso, indicando que los datos faltantes dependen sólo del nivel socioeconómico. En este caso, los datos faltantes son considerados un MAR. La probabilidad que ocurra un dato faltante en el ingreso dependerá del nivel socioeconómico y no del ingreso. Esto es:

$$P\ (Ingreso = Missing/Nivel, Ingresos)\ =\ P\ (Ingreso = Missing/\ Nivel)$$

Para que los datos faltantes en la variable ingreso tenga un mecanismo NMAR, entonces la ocurrencia de los datos faltantes sólo dependerá del propio ingreso. En la columna NMAR, los datos faltantes del ingreso aparecen para cualquier nivel socioeconómico (A, B o C) y para valores del ingreso menores o igual a 60, indicando que los datos faltantes dependen sólo del ingreso. En este caso, los datos faltantes son considerados un NMAR. La probabilidad que ocurra un dato faltante en el ingreso, dependerá del ingreso y no del nivel socioeconómico. Esto es:

$$P\ (Ingreso = Missing/Nivel, Ingresos)\ =\ P\ (Ingreso = Missing/\ Ingresos)$$

Es común referirse a los procesos MAR como mecanismos de no respuesta ignorable, en tanto que MNAR significa que la falta de respuesta no puede ser ignorada en el proceso de construcción del estimador ni al analizar las relaciones de causalidad entre variables.

Existen dos patrones de datos faltantes: univariados y multivariados. El patrón univariado, se presenta cuando sólo existe una variable con datos faltantes. El patrón multivariado, ocurre cuando existen más de dos variables con datos faltantes. A su vez, el patrón multivariado puede tener un patrón monótono o arbitrario. El patrón monótono, ocurre cuando existe una estructura determinística en la ocurrencia de los datos faltantes entre el conjunto de variables. El patrón arbitrario, ocurre cuando los datos faltantes aparecen sin ninguna estructura.

2.1.3 Pruebas para evaluar el mecanismo de datos faltantes

La mayoría de los métodos o técnicas para el manejo de los datos faltantes suponen que dichos datos provienen de un mecanismo completamente aleatorio (MCAR), esto es que los valores perdidos no dependen de las variables con datos observados ni de las variables con datos faltantes (no dependen de la matriz de datos). Según (Rubin, 1976), la definición de un MCAR requiere que los datos faltantes sea una muestra simple aleatoria de un hipotético conjunto de datos completos. Esto implica que los casos con datos faltantes y casos completos provienen de la misma población (el mismo vector de media y matriz de covariancias); a esto se conoce como la condición de homogeneidad de medias y covariancias. Una forma de probar la homogeneidad de medias es formar dos grupos: con datos faltantes y con datos completos, luego se aplica una prueba de hipótesis de diferencia de las medias de los grupos sobre las otras variables en el conjunto de datos. La prueba de homogeneidad de covariancias en forma, se aplicar una prueba de hipótesis de diferentes variancias. Si se encuentra que existe evidencia de un similar vector de medias y matriz de covariancias entre los casos con datos faltantes y completos, entonces el patrón de datos faltantes son un MCAR; si hay diferencias en los dos conjuntos de datos, entonces el patrón de datos faltantes no es MCAR, lo que da la posibilidad que sea un MAR o NMAR. Se plantea dos pruebas para evaluar la existencia del patrón de datos faltantes; una univariada y otra multivariada.

1) Comparaciones univariadas (Pruebas t)

Una forma simple de evaluar si el mecanismo de datos faltantes es MCAR, es usar la estadística "t" para probar la hipótesis de la diferencia de medias independientes entre los grupos de datos faltantes y no faltantes (Dixon, 1983). El método consiste en usar la variable que tiene datos faltantes para separar en dos grupos los datos (faltantes y completos) el resto de variables y luego aplicar pruebas "t" para evaluar si hay diferencias estadísticamente significativas entre las medias en los dos grupos. Existe un MCAR, cuando hay evidencia estadística que las medias con los datos faltantes son similares con los no faltantes. Por lo tanto, si resulta no significativa la prueba "t" de diferencia de medias, entonces hay evidencia de un MCAR en los datos faltantes, y si resulta significativa se sugiere que los datos siguen un MAR o MNAR.

El procedimiento de prueba de hipótesis consiste en suponer dos poblaciones que corresponden a los datos completos y a los datos faltantes, cuyas medias son denotadas por

$\mu_{Completos}$ y $\mu_{Faltantes}$ respectivamente. Esta prueba se realiza para cada una de las variables que no tienen datos faltantes.

- **Las hipótesis formuladas son**:

$$H_0: \mu_{Completos} = \mu_{Faltantes}$$
$$H_1: \mu_{Completo} \neq \mu_{Fatantes}$$

- **Prueba estadística.**

Corresponde a una prueba t para la diferencia de medias independientes, dónde se supone que las variancias poblacionales son desconocidas. Existen dos casos que depende si las variancias son similares (homogéneas) o diferentes (heterogéneas).

Caso 1. Si las variancias poblacionales son similares. La prueba estadística t, será la siguiente:

$$t_c = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\left(\frac{(n_1-1)xS_1^2 + (n_2-1)xS_2^2}{(n_1+n_2-1)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)} = \sout{t_{Tab}}$$

Caso 2. Si las variancias poblacionales son diferentes. La prueba estadística t, será la siguiente:

$$t_c = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}} \sim t_{(g)} = \sout{t_{Tab}} \qquad \text{dónde: } g = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

- **Decisión estadística.**

Se rechaza H_0, si el $|t_c| > t_{Tab}$, en caso contrario no se rechaza H_0. Se espera no rechazar H_0, para afirmar que el mecanismo de los datos faltantes ocurre como un MCAR. Por el contrario, si no se rechaza H_0, entonces el mecanismo de datos faltantes puede ser un MAR o MNAR.

2) Comparaciones multivariadas

Un procedimiento para probar si los datos faltantes siguen un MCAR se logra aplicando la prueba de hipótesis propuesta por (Little J. A., 1988) que se basa en el algoritmo EM (Expectarion-Maximization). Little propuso un estadístico de prueba que sigue una distribución Chi-Cuadrado, donde la hipótesis nula (H_0) establece que los datos faltantes siguen un MCAR. Conforme a la regla de decisión, se debe rechazar H_0 cuando el valor del estadístico de prueba Chi-Cuadrado, es mayor al valor en tablas conforme a un determinado de nivel de significación. Esta prueba puede ser vista como una extensión de la prueba univariada "t" a un método multivariado. La prueba estadística se basa en el método de Máxima Verosimilitud, definida como una suma ponderada de diferencias estandarizadas entre los grupos de medias y la media general. El procedimiento de prueba de hipótesis es el siguiente:

- **Las hipótesis formuladas son**:

  H_0: El mecanismo de los datos faltantes es un MCAR

  H_1: El mecanismo de los datos faltantes no es un MCAR

- **La prueba estadística.**

  El estadístico de prueba es una Chi-cuadrado, definida por la siguiente expresión:

$$\chi_c^2 = d^2 = \sum_{J=1}^{J} n_j \left(\hat{\mu}_j - \hat{\mu}_j^{ML}\right)^T \hat{\Sigma}_j^{-1} \left(\hat{\mu}_j - \hat{\mu}_j^{ML}\right)$$

  Dónde:

n_j Es el número de observaciones con datos faltantes en el patrón j

$\hat{\mu}_j$ Contiene la media de la variable para los casos con datos faltantes en el patrón j

$\hat{\mu}_j^{ML}$ Contiene el estimador de MV de la media general

$\hat{\Sigma}_j^{-1}$ Es el estimador de MV de la matriz de covariancias

$d^2 \sim \chi^2_{\sum k_j - k}$, donde k_j es el número de variables completas para el patrón j y k es el número total de variables.

- **Decisión estadística.**

  Se rechaza H_0, si el $\chi_c^2 > \chi_{Tab}^2$, entonces el mecanismo de datos faltantes puede ser un MAR o MNAR; en caso contrario si no se rechaza H_0 se puede afirmar que el mecanismo de los datos faltantes ocurre como un MCAR. Por lo tanto, se espera no rechazar H_0.

2.1.4 Datos faltantes en la Encuesta Nacional de Hogares – ENAHO

En el Perú la Encuesta Nacional de Hogares (ENAHO) es el instrumento utilizado por el Instituto Nacional de Estadística e Informática (INEI) para recopilar datos con la finalidad de generar indicadores que permitan conocer el estado y la evolución de la pobreza, el bienestar y las condiciones de vida de la población, así como para efectuar diagnósticos y medir el alcance de los programas sociales (alimentarios y no alimentarios) en la mejora de las condiciones de vida de la población peruana. La ENAHO es una fuente de información útil para instituciones públicas y privadas, y proveedora de datos para la realización de muchas investigaciones (INEI, 2015), pero se enfrenta al problema de la no respuesta, principalmente en las variables de naturaleza cuantitativa como los ingresos. Por lo cual es necesario aplicar algún tipo de estimación de los ingresos no declarados, con el fin de considerar en el análisis la mayor cantidad posible de casos. Según las cifras reportadas en la ficha técnica de la encuesta ENAHO sobre condiciones de Vida y Pobreza 2016, se registró el año 2015 un índice de no respuesta a nivel nacional de más de 30% en los estratos Socioeconómicos “A” y “B”. Esto implica que los diversos tipos de estudios que se basan en estos datos (distribución del ingreso, pobreza e indigencia, evolución de ingresos de la población, evolución de ingresos sectoriales, estrategias familiares, etc.) se ven limitados, sesgados o expresan en forma parcial la información que se pretende analizar. El problema de la no respuesta en las encuestas puede ser resuelto con la aplicación de métodos de imputación como se muestra en la revisión bibliográfica.

En (Lindenboim, Graña, J., & Kennedy, D., 2006), menciona que la no certeza de sub-declaración en perceptores de ingresos fijos y la limitada incidencia del volumen de ingresos de los posibles sub declarantes (rentas, ganancias empresariales y trabajadores por su cuenta), se puede considerar como una no declaración de ingresos. Según (Salvia & Donza, E., 1999), la no respuesta o respuesta parcial, puede generar serios impedimentos al realizar los análisis: debido a estos “casos perdidos” los estudios sobre remuneraciones o ingresos familiares están impedidos de hacer inferencias al total de la población por el recorte que sufre la muestra.

Asimismo, los análisis de asociación también se ven afectados, a no ser que se asuma a ciegas el supuesto por demás riesgoso que los casos perdidos presenten distribuciones multidimensionales semejantes a los registros con ingresos informados. Esta afirmación se refuerza por el hecho que en los estudios de distribución del ingreso basados en hogares y los estudios de pobreza, generalmente, la no declaración o declaración parcial de ingresos de un perceptor del hogar impide la consideración de la totalidad de los componentes del hogar en el estudio. Induciendo, además, en segunda instancia, posibles alteraciones en la aplicación de series temporales en las cuales no se podrán diferenciar el efecto generado por el cambio del perfil de perceptores, factores contextuales o cambios metodológicos en el proceso de medición.

2.1.5 Tratamiento de los datos faltantes

La falta de repuesta trae como consecuencia resultados deficientes e incluso inválidos que puede llevar a una pérdida de toda la investigación, distorsión de las frecuencias marginales y/o conjuntas de las variables, sesgos en las estimaciones, disminución del tamaño de la muestra y todo lo que esto implica (aumento del error de muestreo, falta de representación en grupos o variables, estimaciones imposibles de obtener).

Según (García-Laencina, P., Sancho-Gómez, J., Figueiras-Vidal, A., & Verleysen, M., 2009), el manejo de datos faltantes puede tratarse desde cinco diferentes tipos de enfoques:

i) **Análisis de datos completos.** Este procedimiento consiste en eliminar todos los datos faltantes y sólo considerar los completos en el análisis, es conocido como “Listwise deletion” o “Eliminación por lista” y se basa en que el patrón de los datos faltantes es un MCAR (Completamente al azar). Este procedimiento puede justificarse, cuando exista una gran cantidad de datos completos disponibles. Su ventaja es la forma más sencilla y práctica para seguir con el análisis estadístico, considerando el mismo tamaño de muestra para todas las variables. La principal desventaja, es la pérdida de eficiencia y aparición de un sesgo en los estimadores al disminuir el tamaño de muestra (subestimando la variabilidad) debido a la eliminación de las observaciones incompletas en todas las variables, lo que implica un posible cambio en la distribución de los datos.

ii) **Análisis de datos disponibles.** Se utilizan todas las observaciones que tienen valores observados para las variables. Es conocido como “Pairwise deletion” o “Eliminación por pares” y se basa que el patrón de los datos faltantes es MCAR (Completamente al azar).

Su principal ventaja es que se utiliza todos los datos para las variables que se analizarán. La desventaja es la ocurrencia de errores en la estimación, porque que el número de observaciones para cada una de las variables es diferente. Es así, que las covariancias y correlaciones se calcular con pares de datos disponibles. Las matrices de covariancias y correlaciones pueden ser definidas no positivas.

iii) **Imputación de datos faltantes.** El uso de los métodos de imputación que tienen como finalidad completar los datos faltantes reemplazándolos por valores estimados a partir de los datos completos. Existen una variedad de métodos de imputación que se basan en buscar y evaluar las relaciones existen entre los datos completos y los faltantes en la base de datos para aplicar algún método o técnica para la estimación de los datos faltantes.

iv) **Uso de modelos.** Este método es asumir algún modelo para los datos de entrada el método de máxima verosimilitud para obtener una estimación del modelo. El método de MV que usa como variante de el algoritmo de Máximo-Esperado (EM) que puede manejar los valores faltantes para estimar los parámetros del modelo.

v) **Uso máquinas de aprendizaje.** Se usa técnicas de minería de datos dentro de las máquinas de aprendizaje para problemas de clasificación, que permiten dentro del proceso de aprendizaje estimar los datos faltantes; por ejemplo, los árboles de clasificación.

2.1.6 Métodos de imputación para datos faltantes

En las últimas décadas se ha desarrollado muchos técnicas y métodos con la finalidad de atacar el problema de la no respuesta, tratando de mejorar las propiedades estadísticas de las opciones tradicionales que son de fácil aplicación, entre los que se encuentran la eliminación de datos (listwise) y el pareo de observaciones (pairwise). La aplicación de estos métodos significaba trabajar únicamente con las observaciones que disponen de los datos completos para todas las variables, este procedimiento tiene el inconveniente a la posibilidad de una gran pérdida de datos, lo que motivó al surgimiento de nuevos métodos que no presenten este problema, es así que se da origen a los métodos de imputación de datos. Se entiende por imputación al proceso de estimar y rellenar valores faltantes usando los datos disponibles, con el objetivo de obtener un conjunto de datos completos y consistentes que puedan ser analizados posteriormente con las técnicas estadísticas.

Los nuevos aportes en la imputación de datos se realizaron en el año 1932 por Wilks, quien utilizó la sustitución de datos faltantes apoyándose en la media de las variables completas. Posteriormente y ante los reportes de introducción de sesgos por parte de esta metodología, es que en la década de los sesenta se originó un interés por la búsqueda de correcciones y de nuevos métodos, en 1960, se planteó la imputación por regresión Buck (1960), cuya idea básica era el reemplazo o sustitución de los datos faltantes por valores predictivos de una ecuación de regresión, este método también presentaba inconvenientes con respecto a la introducción de sesgos en las varianzas y covarianzas. En 1976 Rubin propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística. Posteriormente, la aparición del algoritmo Expectation Maximization (EM) permitió generar estimadores robustos a partir de la aplicación de la estimación por máxima verosimilitud en donde las observaciones faltantes se asumen como variables aleatorias y los datos imputados se generan por el método de EMV (Dempster, Laird, N. M., & Rubin, D. B., 1977). En la década de los 90, (Todeschini, 1990) propuso el k-vecino más cercano como método para la estimación de los valores perdidos. Una taxonomía para clasificar los diversos métodos y técnicas de imputación de datos faltantes puede verse en (Goicoechea, 2002) Las técnicas de imputación se pueden clasificar de la siguiente manera:

Técnicas fundamentadas en información externa

Cuando son basadas en variables relacionadas con una encuesta perteneciente a otras bases de datos o reglas previas. Entre estas se encuentran:

a) **Métodos deductivos.** Cuando los datos faltantes se deducen con cierto grado de certidumbre de otros registros completos del mismo caso, siguiendo algunas reglas específicas.
b) **Tablas Look-up.** Cuando se hace uso de una tabla con información relacionada, como fuente de datos externa para imputar los datos faltantes.

Técnicas determinísticas

Cuando al repetir la imputación en varias unidades bajo las mismas condiciones, producirá las mismas respuestas.

a) **Imputación de la media, mediana o moda.** Se llena el vacío del dato faltante de cada variable con la media de los registros no faltantes en caso de variables cuantitativas, o con

la moda en caso de variables cualitativas. Tiene como desventaja la modificación de la distribución de la variable haciéndose más estrecha ya que reduce su varianza, además, no conserva la relación entre variables y se debe asumir una MAR. Su ventaja es la facilidad de la aplicación del método.

b) **Imputación de media de clases.** Las respuestas de cada variable son agrupadas en clases disjuntas con diferentes medias, y a cada registro faltante se le imputará con la media respectiva de su grupo. Tiene las mismas desventajas que el caso anterior, pero en menor proporción por estar agrupadas. Igualmente es de fácil aplicación.

c) **Imputación por regresión.** Se ajusta un modelo lineal que describa a y, variable a imputar, para un conjunto X de variables auxiliares que se deben disponer. Resuelve el problema de la distorsión de la distribución de la variable a imputar, pero puede crear inconsistencias dentro de la base de datos, pues podría obtenerse valores "imposibles", ya que el valor y es obtenido de variables auxiliares.

d) **Emparejamiento media.** Se lleva a cabo el método (e) donde el valor de y (estimado) es comparado con casos completos, y el caso más cercano correspondiente provee el valor imputado.

e) **Imputación por el vecino más cercano.** Se identifica la distancia entre la variable a imputar y, y cada una de las unidades restantes (x o variables auxiliares) mediante alguna medida de distancia, entonces se determina la unidad más cercana a y, usando el valor de esta unidad cercana para imputar el faltante.

f) **Algoritmo EM (Expectation Maximization).** Se basa en la función de máxima verosimilitud, permite obtener estimaciones máximo-verosímiles (MV) de los parámetros cuando hay datos incompletos con unas estructuras determinadas. Resuelve de forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos en cada iteración (Little y Rubin, 2002). Este algoritmo tiene la ventaja de que puede resolver un amplio rango de problemas, incluyendo problemas no usuales que surgen de la pérdida o datos incompletos, como lo es la estimación de los componentes de la varianza.

g) **Redes Neuronales.** Son sistemas de información procesados, que reconocen patrones de los datos sin algún valor perdido para aplicarlo a la data a imputar. Estas redes son más usadas para variables cualitativas que cuantitativas, siendo más adecuadas cuando la distribución es no lineal. No es aconsejable cuando hay registros atípicos que distorsionan

la red. Son costosos y requieren de capacitación del analista, así como de "software" adecuado.

h) **Modelos de series de tiempo.** Se asume que la data perdida ocurre de tal forma, y en tal sistema, que el problema se reduce a una situación, en la cual, hay una serie de tiempo, donde una(s) serie(s) de observaciones están perdidas, haciendo óptimo el uso de interrelaciones entre sucesivas observaciones en cada serie de tiempo, mediante el uso de un modelo adecuado para estas series.

Técnicas aleatorias o estocásticas

Son aquellas que cuando se repite el método de imputación bajo las mismas condiciones para una unidad, producen resultados diferentes.

a) **Imputación aleatoria de un caso seleccionado.** Para cada caso con una celda faltante, se selecciona un donante aleatoriamente para ser asignado al dato faltante.

b) **Imputación aleatoria de un caso seleccionado entre clases.** Se realiza de igual forma que para el caso (a) pero se lleva a cabo dentro de clases previamente definidas.

c) **Imputación secuencial Hot-Deck.** Cada caso es procesado secuencialmente. Si el primer caso tiene un dato faltante, este es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa. Si el valor no está perdido, éste será el valor inicial y es usado para imputar el subsiguiente dato faltante. Entre las desventajas se encuentra que cuando el primer registro está perdido, se necesita de un valor inicial, (generalmente obtenido de manera aleatoria), además cuando se necesitan imputar muchos registros se tiende a emplear el mismo registro donante, llevando esto a su vez la pérdida de precisión en las estimaciones.

d) **Imputación jerárquica Hot-Deck.** Similar al método secuencial anterior. En esta se organizan dentro de clases haciendo uso de variables auxiliares en forma de una estructura jerárquica. Si el donante no es encontrado en un nivel de clasificación, las clases pueden ser colapsadas en grupos más anchos hasta que el donante sea encontrado.

e) **Imputación por regresión aleatoria.** Se hace primero un procedimiento de regresión, luego un término residual es adicionado para imputar los valores de y. Este término de error puede ser obtenido de diferentes maneras, una de ellas es a través de los residuos del

modelo de regresión, generado con registros completos, eligiendo uno de estos residuos aleatoriamente.

f) **Imputación por regresión logística.** Similar a la técnica anterior, pero para imputar variables cualitativas.

En (Restrepo Estrada & Marín Diazaraque, 2012), se presenta el problema del manejo de encuestas con datos faltantes y se evalúan algunos métodos de imputación en las variables de ingresos y ganancias de los ocupados (asalariados y cuenta propia) en la Gran Encuesta Integrada de Hogares (GEIH) de Colombia de 2010. En el estudio se evaluaron siete métodos de imputación analizando para el total de la muestra y por grupos de estratos de la vivienda: la eliminación de casos, la imputación por media no condicionada, imputación por regresión estocástica, el hot-deck, el hot-deck con regresión, la imputación múltiple normal multivariada y la imputación múltiple con ecuaciones encadenadas. Se concluye que al no contar con porcentajes altos de no respuesta y dado que es posible que los datos faltantes sigan un patrón que pueda ignorarse, los resultados de los métodos aplicados son relativamente similares.

En (Medina & Galván, 2007), se evalúa la sensibilidad de los indicadores de pobreza y desigualdad a distintos procedimientos de sustitución de datos faltantes en las variables de Ingresos (sueldos o salarios). Los datos utilizados provienen de la Encuesta Permanente de Hogares (EPH), realizada por el Instituto Nacional de Estadística y Censos de Argentina (INDEC) en el 2004. Para la sustitución de los datos faltantes se aplicaron ocho procedimientos de imputación: listwise, medias condicionadas, hot-deck, hot-deck con regresión, regresión condicionada, máxima verosimilitud, imputación simple y dos algoritmos de imputación múltiple. Los resultados del estudio concluyen que cada situación es diferente y la elección del procedimiento de sustitución de datos depende de la variable de estudio, del porcentaje de datos faltantes, del tipo de encuesta que se analice y del uso que se hará de la información imputada.

En (Donza, 2013), se plantea la necesidad de realizar una imputación de los ingresos no declarados en la Encuesta Permanente de Hogares (EPH) del Instituto Nacional de Estadísticas y Censos (INDEC) para el aglomerado Gran Buenos Aires (GBA) entre los años 1990 y 2010. Con el fin de determinar la incidencia de la no respuesta a las preguntas de ingresos, se evalúa el efecto que esto genera y se recomienda el mejor método de imputación como solución a la problemática de datos faltantes. Para el desarrollo del estudio se aplicaron

los métodos de imputación por la media, imputación deductiva, imputación cold deck, imputación hot-deck, imputación por regresión, imputación mediante el método de regresión secuencial multivariante, estimación por máxima verosimilitud (MV). Luego del análisis de las técnicas más utilizadas en el ámbito académico y profesional, se identificó el procedimiento de máxima verosimilitud como uno de los más eficientes para enmendar la no respuesta.

La imputación en las encuestas es común, debido al hecho de que las encuestas a menudo se enfrentan al problema de la falta de datos. En (De Leeuw, 2001), se describe el problema de los datos faltantes en las encuestas y da sugerencias sobre cómo tratarlos. En (Downey & King, 1998), se evalúan dos métodos para imputar datos de tipo Likert, que a menudo se utilizan en encuestas. Sus resultados muestran que ambos métodos, la media del elemento y la sustitución de la media de la persona, funcionan bien si la proporción de datos faltantes es inferior al 20%. En (Raaijmakers, 1999), se presenta un método de imputación, la sustitución media relativa, para imputar los datos de Likert en encuestas a gran escala. Al comparar el método con otros, concluye que parece ser beneficioso en este contexto. También sugiere que es de mayor importancia estudiar el efecto de la imputación en diferentes tipos de datos y estrategias de investigación que estudiar la efectividad de diferentes estadísticas. En (Chen & Shao, 2000), se evalúan la imputación de k-NN con k = 1 para los datos de la encuesta, y muestran que el método tiene un buen desempeño con respecto al sesgo y la variancia de la media de los valores estimados.

En (Schmitt, Mendel & Guedy, 2015), se compara 6 métodos de imputación: media, vecinos K más cercanos (KNN), medias K difusas (FKM), valor singular descomposición (SVD), análisis de componentes principales bayesianos (bPCA) e imputaciones múltiples por ecuaciones encadenadas (RATONES). La comparación se realizó en cuatro conjuntos de datos reales de varios tamaños (de 4 a 65 variables), bajo una falta completamente al azar y basado en cuatro criterios de evaluación: error cuadrático medio de raíz (RMSE), error de clasificación no supervisado (UCE), error de clasificación supervisado (SCE) y tiempo de ejecución. Los resultados sugieren que bPCA y FKM son dos métodos de imputación de interés que merecen mayor consideración en la práctica.

2.2 Método de imputación de datos con k-vecino más cercano

La técnica del k vecino más cercano (**k-NN**: k nearest neighbors), forma parte de las técnicas de aprendizaje automático basado en ejemplos y se usa en muchas aplicaciones de minería de datos para el tipo de aprendizaje supervisado en las tareas de predicción o clasificación y no supervisadas en el agrupamiento de observaciones. El algoritmo más sencillo para la búsqueda del vecino más cercano es el conocido como fuerza bruta o exhaustivo, que calcula todas las distancias de un individuo a los individuos de la "muestra de entrenamiento" (individuos donantes) y asigna al conjunto de vecinos más cercanos cuya distancia sea la menor. En la actualidad se han desarrollado algoritmos eficientes que evitan recorrer exhaustivamente todo el conjunto de entrenamiento.

Entre los métodos de imputación Hot-deck, la técnica del k vecino más cercano se está utilizando con mayor frecuencia como método para imputar datos de una o más variables que presenta valores faltantes. El método de imputación k-NN se aplica a los datos completos para realizar el proceso de imputación. Las observaciones con datos faltantes en las Y, se denotan como el conjunto destino (variables destino) y las observaciones con datos completos en las X's se denomina como conjunto donante (variables donantes). El procedimiento de imputación se basa en seleccionar un conjunto de variables donantes (X) que tienen datos completos y que estén correlacionadas con las variables destino (Y). El método de imputación k-NN, consiste en reemplazar los valores faltantes de una o más variables destino Y usando un conjunto de variables donantes X's con valores conocidos para todas las observaciones, para lo cual se usa una métrica de distancia que permite medir la cercanía entre las observaciones Y y X, para llevar el proceso de reemplazar los valores faltantes en Y con los k-NN más cercanos de los valores de X.

En (Batista & Monard, M.C., Imputación, 2002), se analiza el rendimiento del k- vecino más cercano como un método de imputación, comparado con el rendimiento obtenido por el método de imputación de la media o la moda, y por los algoritmos C4.5 y CN2. Los estudios se realizaron utilizando cuatro bases de datos del repositorio de UCI: Bupa, CMC, Pima y Breast; considerando la imputación con diferentes porcentajes de datos faltantes, número de atributos cualitativos y cuantitativos y valores de k en las bases de datos. El método del 10-KK vecino más cercano mostró mejores resultados, incluso para conjuntos de entrenamiento que tienen una gran cantidad de datos faltantes.

En (Jonsson & Wohlin, 2006), menciona que es común encontrar datos faltantes en encuestas, lo cual trae consecuencias en el análisis estadístico con resultados sesgados. Propone tres etapas para el proceso de imputación: remoción, imputación y evaluación. La etapa de remoción permite generar artificialmente un conjunto de datos incompletos a partir de los completos con la finalidad de definir el tipo de patrón de datos faltantes MCAR o MAR. Este conjunto generado es usado para la imputación del k-NN y la identificación del valor de k. En esta investigación, se aplica el k vecino más cercano como método de imputación con datos faltantes con datos medidos en escala de likert en un contexto de ingeniería de software. Se comparan y evalúan cuatro métodos de imputación: sustitución aleatoria por sorteos, imputación aleatoria, imputación por la mediana y por la moda. Se considera la evaluación del k-NN con diferentes valores de k, proporciones de datos faltantes, estrategias de selección de vecinos y números de atributos. Los resultados muestran que el método k-NN funciona bien, incluso cuando faltan muchos datos, pero tiene fuerte competencia de la imputación por la mediana y la imputación por la moda. Sin embargo, a diferencia de estos métodos, k-NN tiene un mejor rendimiento con muchos atributos de datos. Se sugiere que un valor adecuado de k es aproximadamente la raíz cuadrada del número de casos completos, aumenta la capacidad de imputación del método. En (Song, Shepperd, & Cartwright, 2005), se evalúan las diferencias entre los mecanismos de datos faltantes el MCAR y el MAR usando como métodos de imputación la técnica del k-NN y la imputación de la media por clase, bajo un contexto en la predicción del esfuerzo del proyecto de software. Los resultados del estudio indicaron que no existen diferencias notorias en el tipo de patrón de datos faltantes y los métodos de imputación. Sin embargo, la imputación de la media por clase obtuvo un rendimiento ligeramente mejor que el K-NN.

En (Zainuri, Jemain, & Mura, 2015), en un estudio sobre la calidad del aire en varias estaciones de Malasia, se presenta el problema de datos faltantes (missing). Se aplican diferentes métodos de imputación y se comparan utilizando los datos de las estaciones de Malasia. Los datos faltantes para varios casos se simulan aleatoriamente con 5, 10, 15, 20, 25 y 30% faltantes. Seis métodos utilizados en este documento fueron la sustitución de la media y la mediana, el método de maximización de expectativa (EM), la descomposición de valores singulares (SVD), el método K-vecino más cercano (KNN) y el método secuencial K-vecino más cercano (SKNN). El rendimiento de las imputaciones se compara utilizando el indicador de rendimiento: el coeficiente de correlación (R), el índice de acuerdo (d) y el error absoluto medio (MAE). Con base en el resultado obtenido, se puede concluir que EM, KNN y SKNN

son los tres mejores métodos. Se obtienen los mismos resultados para las ocho estaciones de monitoreo utilizadas en este estudio

Según (Eskelson, Tenmesgen, Lemay, Barret, & Crookston, 2009), en el inventario forestal las bases de datos de monitoreo casi siempre están incompletas, desde datos faltantes por solo unos pocos registros o algunas variables sobre todo al tomar información para grandes extensiones de tierra. En una amplia variedad de situaciones los métodos de imputación k-NN han sido aplicados para completar observaciones en variables que faltan en algunos registros (variables Y), utilizando variables relacionadas que son disponible para todos los registros (variables X). Se realiza una revisión de las ventajas y debilidades de la imputación de k-NN. Se encuentra que una ventaja del k-NN en la imputación de datos faltantes en el inventario forestal, es la posibilidad de controlar la variabilidad de las zonas forestales y seleccionar diferentes conjuntos de X-variables para mejorar las propiedades estadísticas de los estimadores.

En (Tutz & Ramzan, 2014), se realiza un estudio para evaluar la eficiencia del k-NN para la imputación de valores perdidos. Los resultados de la simulación muestran que la estimación de imputación ponderada propuesta funciona mejor que la fija o la de enfoque no ponderado. También se comparan las métricas de distancia L1 y L2 en la ponderación para imputación de datos faltantes y se encontró que la métrica L2 es ligeramente mejor que L1 métrico. El uso de funciones de kernel para el cálculo de pesos hace disminuir el error de imputación. Los resultados de la simulación sugieren que el kernel gaussiano proporciona MSEs más pequeñas que otras funciones de kernel. Para hacer frente al problema de los datos de alta dimensión, se propone una imputación con una selección ponderada de predictores. El procedimiento utiliza validación cruzada para una selección óptima de los parámetros de ajuste. En particular, para datos altamente correlacionados, el procedimiento de imputación de k-NN propuesto da mejores resultados, también en el caso de una alta proporción de datos faltantes.

2.2.1 El método k-vecino más cercano

El método del k vecino más cercano, puede considerarse como una técnica no paramétrica que no necesita el conocimiento de una distribución de probabilidad de los datos. El k vecino más cercano se usa en el aprendizaje supervisado o no supervisado dentro de las técnicas de aprendizaje de máquina. El K-NN es un algoritmo que se basa en instancias (observaciones), esto es, que no se desarrolla explícitamente un modelo, sino, que memoriza las instancias del conjunto de entrenamiento que son usadas en la clasificación. El k-NN, en el aprendizaje

supervisado permite clasificar una nueva observación a alguna de las clases establecidas, a partir de evaluar sus cercanías a cada una de las observaciones (conjunto de entrenamiento) a través de usar alguna métrica de distancia. El algoritmo del k-NN, se basa en clasificar una observación nueva (patrón) en la clase más frecuente a la que pertenecen sus k vecinos más cercanos. La fase de entrenamiento del algoritmo consiste en almacenar los valores de las variables predictivas y de la variable objetivo sus etiquetas de las clases definidas. En la fase de clasificación, la evaluación de la observación que se desea clasificar (patrón) a la que no se conoce su clase es representada por un vector en el espacio característico. Se calcula la distancia entre las variables almacenadas y la nueva observación, y se seleccionan los k ejemplos más cercanos. La nueva observación es clasificada con la clase que más se repite (más frecuente) en los vectores seleccionados.

El k-NN es muy eficiente para el problema de clasificación de observaciones. En la tarea de clasificación de un patrón X, la regla se basa en la vecindad (distancia) sobre la búsqueda de un conjunto de observaciones que corresponde a los k vecinos más cercanos a dicho patrón a ser clasificado. Luego, para la asignación del patrón a clasificar se aplica una regla de decisión.

Reglas para asignar el patrón X a alguna de las clases:

- **Regla k-NN.** Esta regla se basa en clasificar el patrón X, en la clase más frecuente a la que pertenece al grupo de sus k vecinos más cercanos. Si $K_j(X)$ denota el número de observaciones que pertenecen a la clase $\mathbf{C_j}$ (para j=1, 2,….,m clases) presentes en los k vecinos más cercanos al patrón X al evaluar a través una medida de distancia $d(X,X_i)$, entonces la regla de clasificación puede expresarse:

$$X \Longrightarrow se\ asigna\ a\ C_h, \quad si:\ K_h(X) = \max_{j=1,2,\ldots,m}\{K_j(X)\}$$

$$donde:\ k = \sum_{j=1}^{m} K_j(X)$$

La decisión para la clasificación será, etiquetar el patrón X con la clase mayoritaria $\mathbf{C_h}$ **(**el mayor valor de los K**j**(X). Si existiera empate entre las clases, entonces se clasifica el patrón al azar en alguna de las clases.

- **Regla (k ,t)-NN.** Se trata que la clase mayoritaria en los k vecinos más cercanos tenga un número de representación mayor que un umbral t.

$$X \Longrightarrow se\ asigna\ a\ C_h, \quad si:\ K_h(X) = \max_{j=1,2,\dots,m}\{K_j(X)\} \geq t$$

- **Regla (k, tc)-NN.** Para especificar un umbral diferente para cada clase (t_c). Esto permite controlar el "grado de confianza" para aceptar una clasificación en determinadas clases críticas.

$$X \Longrightarrow se\ asigna\ a\ C_h, \quad si:\ K_h(X) = \max_{j=1,2,\dots,m}\{K_j(X)\} \geq t_c$$

<u>Criterios para aplicar a las distancias:</u>

- **K-NN con distancia media.** Se asigna el nuevo patrón X a la clase cuya distancia media sea la menor.
- **K-NN con distancia mínima.** Se inicia seleccionando un caso por clase, normalmente el caso más cercano al baricentro de todos los elementos de dicha clase. En este paso se reduce la dimensión del fichero de casos a almacenar de *N* a *m*. A continuación, se asigna el nuevo caso a la clase cuyo representante esté más cercano. Este procedimiento puede verse como un 1-NN aplicado a un conjunto de *m* casos (uno por cada clase).
- **K-NN con ponderación.** Consiste en que el k-NN aplique una ponderación de las observaciones seleccionados en que los *K* casos no se contabilicen de igual forma, sino que se tenga un valor de la distancia de cada caso al nuevo caso que se pretende seleccionar.

Para clasificar el patrón X aplicando el k vecino más cercano, se considera los siguientes pasos:

1. Se selecciona alguna métrica de distancia.
   - Se calcular las distancias del patrón (X) a clasificar con cada una de las observaciones del conjunto de entrenamiento: $D(X,X_i)$
   - Si es necesario se debe estandarizar las variables
2. Se ordenan las distancias de menor a mayor (D_i).
3. Se determinan las k distancias menores al patrón P_0 y se etiqueta con la clase C_j que le corresponde.

$D_1 = D(X, X_i)$, Cj, para algún j=1,…,m

$D_2 = D(X, X_i)$, Cj, para algún j=1,…,m

. . .

$D_k = D(X, X_i)$, Cj, para algún j=1,…,m

4. Se calcula los $K_j(X)$, el número de observaciones que pertenecen a cada una de las clases C_j en los k vecinos más cercanos.

5. Luego se asignar el patrón P_0 a la clase mayoritaria (C_h), el mayor valor de los $K_j(X)$. Si hubiera empate, se elige al azar la clase. Aplicando la regla de asignación:

$$X \Longrightarrow se\ asigna\ a\ C_h, \quad si:\ K_h(X) = \max_{i=1,2,\ldots,m} \{K_i(X)\}$$

Ejemplo. Suponga que se desea clasificar por el método k vecino más cercano: 3-NN considerando que existen dos clases (Clase1 y Clase 2). En la figura 1, presenta la distribución de las distancias de los ejemplos de cada clase con el patrón (X). Se observa que al considerar los 3 vecinos más cercanos al patrón X, los corresponden valores de los $K_i(X)$ para la Clase 1 y Clase 2 son: $K_1(X)=1$ y $K_2(X)=2$. Por lo tanto, el patrón X se clasifica a la Clase 2 por tener el mayor valor $K_i(X)$.

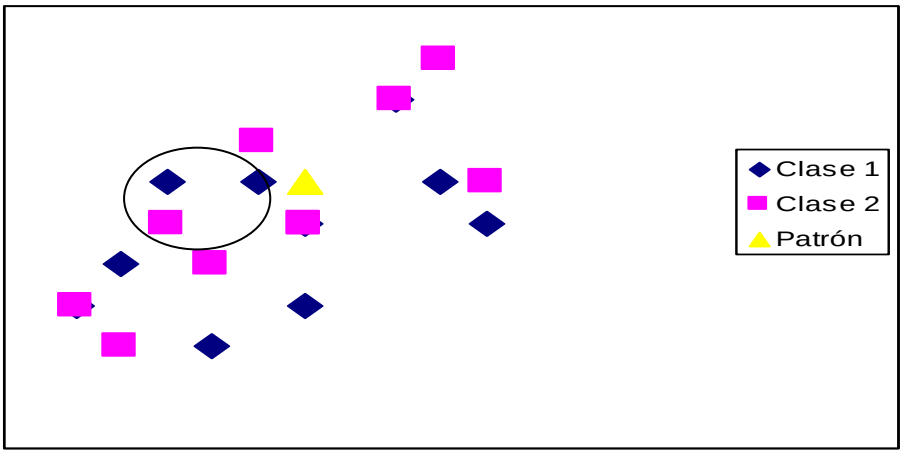

Figura 1. Ejemplo de aplicación del k-NN

Ejemplo. Se cuenta con información de los gastos para una muestra de 12 hogares, los cuales están clasificados por tres niveles socioeconómicos. Aplicar el algoritmo k vecino más cercano para valores de k iguales a 3, 5 y 7, para clasificar el hogar P_0. Los datos se presentan en el Cuadro 2.

Cuadro 2. Gastos de hogares clasificados por nivel socioeconómico (Ejemplo)

Hogar	Gastos en servicios (x1)	Gastos en educación (x2)	Gastos en salud (x3)	Nivel socioeconómico (Ci)
1	10	45	12	Bajo
2	12	55	15	Medio
3	14	60	18	Alto
4	15	65	14	Alto
5	10	40	14	Bajo
6	20	130	65	Medio
7	14	65	45	Bajo
8	12	54	25	Alto
9	16	100	60	Medio
10	16	120	65	Alto
11	18	150	58	Medo
12	12	48	32	Medio

Hogar	Gastos en servicios	Gastos en educación	Gastos en salud
Ho	18	25	10

Elaboración propia.

Número de observaciones: n=12

Número de variables: p=3

Número de clases: m=3 (Niveles socioeconómicos: C_1=Bajo, C_2=Medio y C_3=Alto)

Patrón P_0: x_0 = (18, 25, 10)

1. Se calculan las distancias euclidianas del patrón hogar P_0 (x_0) a cada uno de los otros hogares (x_i).

$$d(x_0, x_1) = \sqrt{\sum_{k=1}^{3} (x_{0k} - x_{ik})^2}$$

$$d(x_0, x_1) = \sqrt{\sum_{k=1}^{p} (18-10)^2 + (25-45)^2 + (10-12)^2} = 21.63$$

$$d(x_0, x_2) = \sqrt{\sum_{k=1}^{p} (18-12)^2 + (25-55)^2 + (10-15)^2} = 31.00$$

$$d(x_0, x_3) = \sqrt{\sum_{k=1}^{p} (18-14)^2 + (25-60)^2 + (10-18)^2} = 36.12$$

$$\cdots$$

$$d(x_0, x_{12}) = \sqrt{\sum_{k=1}^{p} (18-12)^2 + (25-48)^2 + (10-32)^2} = 32.39$$

2. Se ordenan las distancias de menor a mayor (Di).
3. Se determinan las k distancias menores al patrón P_0 y se etiqueta con la clase C_j que le corresponde.

$$Para\ k = 3: \quad d_1 = d(x_0, x_5) = 17.46 \quad (C_1 = Bajo)$$
$$d_2 = d(x_0, x_1) = 21.63 \quad (C_1 = Bajo)$$
$$d_3 = d(x_0, x_2) = 31.00 \quad (C_2 = Medio)$$

4. Se calcula los $K_j(X)$, el número de observaciones que pertenecen a cada una de las clases C_i en los k vecinos más cercanos.

$$K_1(X) = 2, K_2(X) = 1, K_3(X) = 0$$

5. Luego se asignar el patrón P_0 a la clase mayoritaria (C_h). Aplicando la regla de asignación:

$$X \Longrightarrow se\ asigna\ a\ C_h, \quad si:\ K_h(X) = \max_{j=1,2,\ldots,m} \{K_j(X)\}$$

Entonces se tiene:

$$K_h(X) = \max_{i=1,2,\ldots,m} \{K_1(X) = 2, K_2(X) = 1, K_3(X) = 0, \} = K_1(X),$$
$$C_h = C_1(Bajo)$$

Por lo tanto, el hogar P_0 pertenece al nivel socioeconómico Bajo.

En el Cuadro 3, se presentan los cálculos de las distancias usando la métrica Euclidiana respecto al patrón (Hogar P_0) y el resto de los hogares. En la última columna se muestra el orden de las distancias según su magnitud.

Cuadro 3. Cálculos de las distancias Euclidianas (Ejemplo)

Hogar	Gastos en servicios (x1)	Gastos en educación (x2)	Gastos en salud (x3)	Nivel socioeconómico (Ci)	**Distancia a P_0**	Distancias ordenadas (Di)
1	10	45	12	Bajo	21,63	d2
2	12	55	15	Medio	31,00	d3
3	14	60	18	Alto	36,12	d6
4	15	65	14	Alto	40,31	d7
5	10	40	14	Bajo	17,46	d1
6	20	130	65	Medio	118,55	d11
7	14	65	45	Bajo	53,30	d8
8	12	54	25	Alto	33,20	d4
9	16	100	60	Medio	90,16	d9
10	16	120	65	Alto	109,79	d10
11	18	150	58	Medo	133,90	d12
12	12	48	32	Medio	32,39	d5

Elaboración propia.

En el Cuadro 4, se presenta el resume los resultados de la clasificación para el patrón (Hogar P_0) considerando los valores de k=3, 5 y 7. Se muestra que para un 3-NN el hogar se clasificó como un nivel socioeconómico Bajo, para un 5-NN el resultado de la clasificación resultó con un empate entre el nivel socioeconómico Bajo y Medio y para 7-NN el hogar se clasifica con un nivel socioeconómico Alto.

Cuadro 4. Resultados de la clasificación k vecino más cercano (Ejemplo)

Valores de k	**$K_h(X)$**	**Clasificación**
3	$K_1(X)=2$	Bajo
5	$K_1(X)=2$ y $K_2(X)=2$	Bajo/Medio (empate)
7	$K_3(X)=3$	Alto

2.2.2 Métricas para la medición de la distancia

La medición de la distancia se relacionada con la noción de proximidad o similitud entre un conjunto de individuos. El método de imputación usa los valores de las variables medidas para todas las unidades observadas X (conjunto referencial) para guiar la imputación de valores Y (conjunto destino) que están medidos solamente para algún subconjunto de la muestra de las unidades observadas, de tal manera que se tienen valores faltantes. Sean X_i a Y_i conjunto de vectores de atributos correspondiente a la i-ésima unidad observada. La imputación de k vecino más cercano, selecciona unidades desde el conjunto referencial para servir como sustitutos de los miembros del conjunto destino usando una medida de similaridad basada sobre los valores de X. La selección de una medida particular de similaridad, puede depender de la relación de los valores de Y sobre los valores de X.

Se usa el término de distancia para valorar la función que mide la disimilaridad entre el i-ésimo (Xi) y j-ésimo (Xj) par de unidades observadas. La forma general para definir la distancia es a través de la forma cuadrática:

$$d_{i,j}^2 = (X_i - X_j)W(X_i - X_j)'$$

Donde Xi es un vector (1xp) de X-variables para la i-ésima unidad observada destino, Xj es un vector (1xp) de X-variables para la j-ésima unidad observada referencial, y W es una matriz simétrica (pxp) de ponderaciones. Según la forma que tome la matriz W, se puede generar una serie de medidas de distancia:

- **Distancia Euclidiana.** Es la medida de distancia más común usada para medir la similaridad. La distancia Euclidiana se genera, cuando la matriz de ponderaciones W corresponde a la matriz identidad (I). Se tiene entonces:

$$d_{i,j}^2 = (X_i - X_j)I(X_i - X_j)'.$$

- **Distancia Mahalanobis.** Se genera cuando la matriz de ponderaciones W, es considerada la inversa de la matriz de variancias-covariancias (S). Se usa para estimar los errores de imputación. Se tiene entonces:

$$d_{i,j}^2 = (X_i - X_j)S^{-1}I(X_i - X_j)'.$$

- **Distancia de correlación canónica.** Cuando la matriz de ponderaciones es derivada desde el análisis de correlación canónica, regresión canónica o correspondencia canónica. Se tiene la matriz de vectores canónicos (Γ) y la matriz de correlaciones canónicas (Λ). Se tiene entonces:

$$d_{i,j}^2 = (X_i - X_j)\Gamma\Lambda^2\Gamma^T(X_i - X_j)'.$$

- **Distancia análisis espectral.** Cuando se realiza una transformación de la matriz X a los valores de $x_i/\sqrt{\sum_{i=1}^{p} x_i^2}$, entonces la forma cuadrática de la matriz de ponderaciones corresponde a la imputación con el análisis espectral.

- **Distancias para variables mixtas.** Cuando se tienen datos de las observaciones que se han evaluado tanto variables cualitativas como cuantitativas, se usa una modificación de la la distancia Gower. Para calcular la distancia entre dos observaciones se usa una ponderación que mide la importancia de cada variable. Así, la distancia entre la i-ésima y j-ésima observación se define:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_i \delta_{i,j,k}}{\sum_{k=1}^{p} w_k}$$

Dónde w_i es el peso y $\delta_{i,j,k}$ es la contribución de la k-ésima variable. Dependiendo si el tipo de variable se tiene. Para variables continuas, se tiene la distancia en valor absoluto dividido por el rango total, esto es:

$$\delta_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k}$$

Dónde $x_{i,k}$ y $x_{j,k}$ son los valores de la i-ésima y j-ésima observación para la k-ésima variable, y r_k es el rango de la k-ésima variable.

2.2.3 Determinación del valor de k

Se constata empíricamente que el porcentaje de casos bien clasificados es no monótono con respecto de k, siendo los valores elegidos para k comprendidos entre 3 y 7. Sin embargo, la elección de cuántos vecinos usar y qué peso usar para calcular los valores promedio no está clara, y algunas veces se elige para cumplir un criterio objetivo; por ejemplo, reducir el Error cuadrático medio. Según (Tuominen, S., Fish, S., & Poso, S., 2003), cuanto mayor es el valor de k, se produce un aumento en el promedio de las estimaciones. Por lo tanto, el valor óptimo

de k es una compensación entre la precisión de las estimaciones y la variación retenida en las estimaciones. Sin embargo, a medida que k se incrementa el sesgo (promedio de la diferencia entre los valores observados y predichos) aumentan para los valores extremos de las variables de interés. Usar un vecino probablemente proporcionaría los mejores resultados si hubiera una alta proporción de observaciones con información completa, porque estas observaciones de referencia representarían bien a la población. Por el contrario, si hubiera una baja proporción de observaciones con información completa, usar más de un vecino podría dar mejores resultados porque el promedio proporcionaría una variedad más amplia de las variables de interés (Y). El uso de demasiados vecinos puede dar como resultado una menor variabilidad en los valores imputados que en la que se presenta en la población debido al promedio de los valores (Mc Roberts, Nelson, M. D., & Wendt, D. G., 2002). Muchos de los algoritmos para aplicar el método de los k vecinos más cercanos encajan dentro de un esquema de búsqueda conocida como esquema de aproximación y eliminación. El esquema general de búsqueda por aproximación y eliminación se resume como:

1. De entre los individuos del conjunto de entrenamiento, se selecciona un candidato a vecino más cercano (aproximación).
2. Se calcula su distancia d al individuo en cuestión.
3. Si la distancia es menor que la distancia del vecino más cercano hasta el momento, d_{nn}, se actualiza el vecino más cercano y se eliminan del conjunto de entrenamiento aquellos individuos que no puedan estar dentro de una hiperesfera de radio d y con centro en la muestra, es decir, se eliminan aquellos individuos que no puedan estar más cerca de la muestra que el vecino más cercano actual.
4. Se repiten los pasos anteriores hasta que no queden individuos por seleccionar en el conjunto de entrenamiento, ya sea porque han sido previamente seleccionados o porque han sido eliminados.

Un procedimiento que se usa, es que a partir de los N valores de Y denominado conjunto objetivo, se extrae una muestra aleatoria **n** valores denominado conjunto de referencia (datos completos) y se consideran que los N – n observaciones como datos faltantes. El valor Y imputado para un elemento en el conjunto objetivo es una función conocida de los valores Y y del conjunto de referencia cuyos valores de X asociados son los más cercanos. Entonces para valores de k y usando una métrica de distancia aplicada a los valores de las X's se van seleccionando los valores imputados en Y. Se va evaluando los posibles valores de k, a través

de calcular el error cuadrado medio como una medida de precisión y usando procedimientos de validación cruzada (McRoberts, 2009).

2.2.4 Selección de variables donantes

Un aspecto importante para aplicar la imputación por el método k vecino más cercano, es referente a la selección del conjunto de variables donantes. La selección de las variables donantes X's, depende de la información disponible y de su correlación que exista con las variables destinos Y (LeMay, V. & Temesgen, H., 2005a). Cuando el número de variables donantes se incrementa, no necesariamente se mejora los resultados de las estimaciones con el proceso de estimación; más bien aumenta la complejidad en la técnica del k vecino más cercano. Según (Packale, P. & Maltamo, M., 2007), presenta un algoritmo para una selección heurística de las variables donantes que se basa en minimizar el promedio ponderado del error cuadrado medio, usando una matriz de ponderaciones W.

Un procedimiento práctico, es hallar el coeficiente de correlación de Pearson entre las variables donantes y destinos y probar la significación. Se seleccionarán aquellas variables donantes que resulten significativas con las variables destino. Procedimientos para seleccionar las variables donantes para la imputación por k vecino más cercano, puede involucrar el uso del análisis de regresión. En este caso, se trata de un problema de seleccionar las mejores variables predictoras (variables donantes) que expliquen la variable dependiente Y (variable destino). Los métodos de selección de variables comúnmente usados en el análisis de regresión, son el Stepwise, Fordward y Bacward. También se aplican criterios el AIC, BIC, Mallow, etc.

2.2.5 Evaluación de los métodos de imputación

El proceso de la imputación de datos debe ser considerado parte de las etapas de la investigación, con la finalidad de obtener conclusiones confiables que sustenten la hipótesis de estudio. Es así, que la atención de los estudios sobre el problema de datos faltantes no debería sólo concentrarse en generar estimadores que satisfagan propiedades estadísticas deseables. Las bondades de los procedimientos de imputación no deben valorarse por el sólo hecho de que permiten completar información para ajustar modelos y probar hipótesis. Los criterios para evaluar la pertinencia de un método estadístico fueron establecidos por (Neyman, J. & Pearson, E.S., 1933) y (Neyman, 1937), y guardan relación con el error cuadrático medio (ECM) y no sólo con el sesgo del estimador.

Existen muchas maneras de llevar a cabo la evaluación y validación de la aplicación de los métodos de imputación de datos, siendo la más común el uso de simulación de datos faltantes. Esto consiste en crear datos faltantes (missing) artificialmente a partir de la base de datos con datos completos. Se define una muestra aleatoria de un porcentaje (5% a 15%) de datos faltantes, de tal manera que estos datos faltantes tendrán un mecanismo al menos MCAR, siendo una de las condiciones para la aplicación de muchos métodos de imputación.

Entonces se tiene una base de datos completa y una simulada con datos faltantes. Se aplican los métodos de imputación que se desea evaluar a la base de datos con datos faltantes, consiguiendo como resultado una base de datos imputada. Se realiza la comparación y evaluación considerando las variables de la base de datos completa e imputada que presentaban datos faltantes. Se puede medir la desviación estándar, el error cuadrado medio, el sesgo de las estimaciones, las distribuciones de frecuencias conjuntas y marginales, entre otras magnitudes de los efectos de imputación, mediante el uso de tablas de contingencia, de distribuciones de frecuencia o prueba de hipótesis, etc. Para ello se hará uso del estadístico T para variables continuas y el estadístico Chi-cuadrado para variables categóricas. Goicoechea (2002) propone una serie de medidas deseables, al evaluar las técnicas de imputación:

1. Precisión en la predicción; el valor imputado debe ser lo más cercano al valor verdadero.
2. Precisión en la distribución; mantener en lo posible las distribuciones marginales y conjuntas.
3. Precisión en la estimación; producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.
4. Imputación plausible: valores aceptables al aplicarles el proceso de edición.

Las estadísticas de ajuste comúnmente utilizadas por otros autores se basan en comparar los valores observados con los estimados en el conjunto de datos objetivo simulado y, en particular, a menudo se calcula la diferencia promedio (a menudo llamada sesgo) y el error cuadrático medio (raíz cuadrada de la diferencia cuadrada promedio: RMSE). Para más de una variable de interés, una pequeña diferencia promedio en una variable podría ser compensada por una pequeña diferencia promedio en otra variable. Además, las grandes diferencias negativas y positivas darían como resultado una diferencia promedio de cero. El RMSE da una mejor indicación de los resultados de imputación, porque las diferencias se

cuadran antes de promediar. Para evaluar los resultados de los métodos de imputación se propone usar el RMSE y el sesgo (LeMay & Temesgen, 2004).

1. Se determina para cada variable X la diferencia del valor observado y el imputado, siendo el sesgo la suma promedio de estas diferencias. Así para la j-ésima variable X:

$$Sesgo = \sum_{i=1}^{n} (x_{ij}^{Obv} - x_{ij}^{Imp})/n$$

2. Se define el RMSE para cada variable X, como la raíz cuadrada del promedio de la suma de cuadrado de la diferencia del valor observado y el imputado. Así, para la i-ésima variable X:

$$RMSE_j = \sqrt{\sum_{i=1}^{n} (x_{ij}^{Obv} - x_{ij}^{Imp})^2/n}$$

3. El sesgo para variable Y se calcula como la suma promedio de las diferencias entre el valor observado y el imputado. Así se tiene:

$$Sesgo = \sum_{i=1}^{n} (y_{i}^{Obv} - y_{i}^{Imp})/n$$

4. Se define el RMSE para la Y, como la raíz cuadrada del promedio de la suma de cuadrados de la diferencia del valor observado y el imputado. Así, se tiene:

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_{i}^{Obv} - y_{i}^{Imp})^2/n}$$

2.2.6 Imputación con el k vecino más cercano

La imputación por el k vecino más cercano, forma parte de uno de los métodos de imputación conocido como Hot-Deck. El método Hot-Deck, se basa en reemplazar los datos faltantes con valores de unidades similares. Existen varios criterios para duplicar los valores de las observaciones completas (variables donantes) en las faltantes (variables destino):

- **Imputación Hot-Deck aleatoria.** La selección de los datos completos para reemplazar en los datos faltantes se realiza por un muestreo con reemplazo aleatoria simple o estratificado por grupos homogéneos.

- **Imputación Hot-Deck por k vecino más cercano.** La selección de los datos completos se realiza a través de la técnica del k vecino más cercano.

La imputación del k vecino más cercano, basa su proceso de reemplazar los datos faltantes por los datos completos, usando alguna métrica de distancia para que permita medir la cercanía del dato faltante con el resto de las observaciones completas. El procedimiento de imputación puede describirse:

1) Se particiona el conjunto de datos X en dos conjuntos: X_{Obs}, conjunto de datos observados y X_{Mis}, conjunto de datos faltantes.
2) Para cada una de las observaciones Yi de los X_{Mis} se calculan con alguna métrica las distancias entre Yi y las observaciones Yj de los X_{Obs} (D (Yi, Yj)).
3) Ordenando las distancias de menor a mayor, se escogen las k observaciones de menores distancias (las más cercanas a Yi). Se genera el conjunto X_k que contiene los k vecinos más cercanos a Yi, el cuál contendrá datos faltantes para una o varias variables.
4) Con las observaciones en X_k, se imputan los datos faltantes en Yi. Si la variable a ser imputada es cuantitativa, entonces el valor a ser imputado es el promedio o mediana dentro de X_k (el promedio de los k vecinos más cercanos) y si es cualitativa, se reemplaza el dato faltante por moda (el valor más frecuente dentro de los k vecinos más cercanos).
5) El proceso continúa hasta terminar con todas las observaciones de X_k.

Un inconveniente que presenta este método de imputación es la selección del valor de k. Se proponen varios criterios, usar validación cruzada para evaluar la tabla de clasificación (aprendizaje supervisado), el valor entre 3 a 7, etc. En el caso de un problema de imputación, cuando se el valor de k es pequeño, la estimación se hará sobre una muestra pequeña con el efecto de tener una mayor variancia. Mientras que, si se considera un valor grande de k, puede aparecer sesgo en las estimaciones.

III. MATERIALES Y MÉTODOS

3.1 Materiales

Los materiales y equipos que se usarán para realizar la presente tesis son los siguientes:

1. Una computadora personal Intel® Core ™ i7. CPU 3.5 GHz. RAM de 4.00 GB.

2. Programa estadístico R. Se descarga programa R versión x64 3.5.2**.** Se descarga e instalan:
   - package VIM
   - library("VIM")

   El paquete VIM (Templ, Alfons, Kowarik, & Prantner, 2016) fue desarrollado para explorar y analizar la estructura de los valores missing por medio de la visualización de datos, aplicar una variedad de métodos para la imputación de datos. Permite trabajar con datos mixtos (cualitativos y cuantitativos), grandes conjuntos de datos y el manejo de datos atípicos u outliers. El paquete VIM, incluye dos métodos de imputación simple, Hot-Deck (aleatorio y secuencial) y k vecino más cercano usando una métrica de distancia generalizada para trabajas con datos mixtos y también ofrece medidas de agregación como la media y la mediana.

3. Programa en R con el package "VIM". Se ha desarrollo el programa con las respectivas funciones usando el package de R "VIM" para obtener los resultados de la imputación de datos faltantes con k vecino más cercano y los otros métodos propuestos para los ingresos de los hogares (Ver Anexo 1. Programa desarrollado para la imputación de datos).

4. Una impresora inyectora HP.

3.2 Metodología

3.2.1 Tipo de investigación y formulación de hipótesis

La investigación es no experimental con diseño descriptivo y exploratorio. Esto debido a que se analizan los datos de los ingresos mensuales de los hogares recopilados por la ENAHO correspondiente al tercer trimestre. Adicionalmente se usan los datos de los gastos que

incurren los hogares, con respecto a la alimentación, servicios del educación y salud, transporte y comunicación, etc.

Tipo de investigación.

Investigación descriptiva y exploratoria, se estudia el comportamiento y se estima los datos faltantes de los ingresos de los hogares aplicando un método de imputación, y comparando los resultados a través de verificar las estimaciones encontradas aplicando intervalos de confianza y correlaciones. También se obtienen la predicción de los datos completos (observados e imputados).

Formulación de hipótesis

El método de imputación con el k vecino más cercano mejora la precisión de los estimadores cuando se tienen datos faltantes en los ingresos de los hogares en la ENAHO 2017 tercer trimestre con respecto a los métodos de la media, mediana y Hot-Deck.

3.2.2 Población y Muestra

- **Población.** La población está compuesta por todos de los hogares del Perú (según la ENAHO 2017 – Tercer trimestre).
- **Muestra.** Los hogares de del Perú considerados en la ENAHO 2017 – Tercer Trimestre.

3.2.3 Descripción de las variables

Para la aplicación de los métodos de imputación para estimar los ingresos de los hogares en la ENAHO, cuyas variables poseen su propio código de identificación. En el Cuadro 5, se presenta las respectivas variables que serán consideradas en el estudio.

Cuadro 5. Relación de variables de la ENAHO

Código	Descripción
P117T	Total gasto mensual servicios de la vivienda
P311T1	Gasto total por todos los rubros en escolaridad
P416	Gasto total en atención y servicios de salud
P513T	Horas trabajó la semana pasada, en su ocupación principal
P513A1	Tiempo trabaja en la ocupación principal (años)
P518	Horas trabajó la semana pasada en sus ocupaciones secundarias
P559D	Gasto por desayuno, almuerzo y cena
P560D	Gasto por transporte y teléfono
P524A1	Ingreso total de la ocupación principal
P538A1	Ingreso total de la ocupación secundaria
P544T	Ingresos extraordinarios

P556T1	Ingreso por transferencias corrientes del país
P556T2	Ingreso por transferencias corrientes del extranjero
P557T	Ingreso por rentas de la propiedad
P558T	Ingreso por otros ingresos extraordinarios

3.2.4 Proceso para la imputación de los ingresos de los hogares en la ENAHO

El procedimiento para el análisis de los datos de la ENAHO, se basa el aplicar el método de imputación del k-vecino más cercano con la finalidad de estimar los datos faltantes de la variable ingresos por hogar en la encuesta ENAHO. Se realizaron los siguientes pasos:

Paso 1. Recopilación de datos

En este paso se procedió a recopilar la base de datos de la ENAHO 2017 – tercer trimestre que se encuentra en el Portal del INEI. En la Figura 2, se muestra el portal del INEI para seleccionar y acceder a las bases de datos de la ENAHO 2017-Trimestre 3. Se observa los diferentes módulos (archivos .sav) en los cuales se almacenan los datos por rubros de información.

Figura 2. Acceso a las bases de datos por el Portal del INEI

Se diseñó una base de datos para integrar el conjunto de tablas relacionadas que se encuentran en el formato .sav (Programa SPSS). Los módulos que se fusionaron son:

- Módulo 100. Información de la vivienda y el hogar

- Módulo 300. Información de la salud
- Módulo 400. Información de la educación
- Módulo 500. Información del empleo e ingreso

Como resultado de este paso, se obtuvo la base de datos integrada con las principales variables para llevar a cabo el proceso de imputación de los datos faltantes del ingreso de los hogares. En la Figura 3, se muestra el proceso de la fusión de los cuatro módulos, los campos usados como clave primaria para relacionar las bases de datos y el resultado de la base de datos integrada.

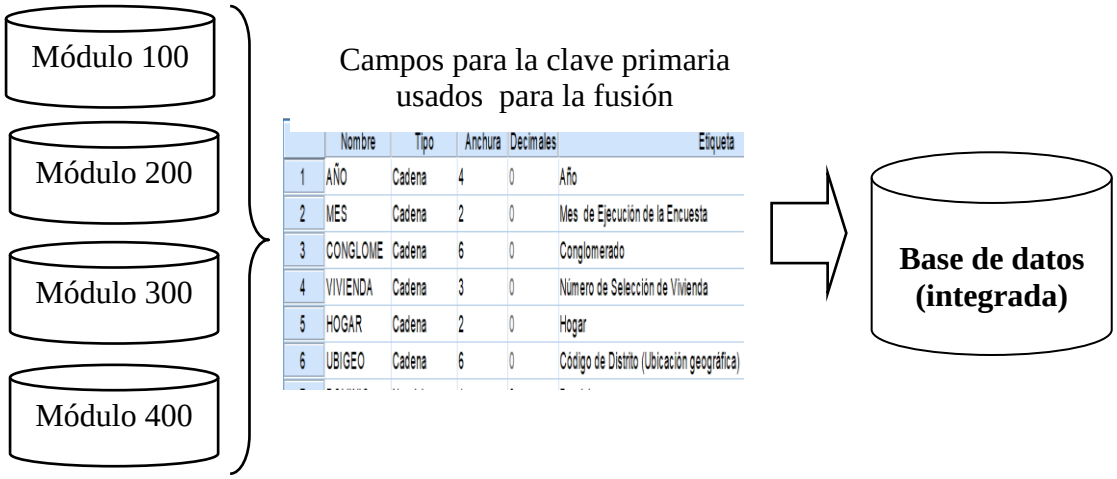

Figura 3. Proceso de fusión de las bases de datos de la ENAHO 2017

Paso 2. Preprocesamiento de datos

En este paso se prepararon y se limpiaron los datos de la base de datos integrada, para lo cual se aplicaron las técnicas para el manejo de datos atípicos, faltantes, inconsistentes y se aplicó transformación de datos; Se obtuvo una base datos consistente para aplicar los métodos de imputación propuestos en esta investigación. Por lo tanto, se realizó:

- **Transformación de variables.** Se procedió a crear las respectivas variables donantes (X's) y la variable destino (Y) para llevar a cabo el proceso de imputación de datos.
- **Manejo de datos faltantes.** Se analizó sólo para las variables X's (donantes) la existencia de datos faltantes. Se procedió a aplicar el procedimiento de Listwise (análisis con datos completos), eliminando los registros (observaciones) de la base de datos en aquellos atributos (variables donantes) que presentaron al menos un dato faltante.
- **Manejo de datos atípicos.** Se elaboraron diagramas de cajas para identificar los datos atípicos u outliers en la variable destino (Y). Se calcularon los cuartiles (P_{25}, P_{50} y P_{75}) y

los límites de seguridad inferior y superior y luego se eliminaron los datos atípicos encontrados.

Límite de seguridad inferior: $LSI = P_{25} - 1.5 \times (P_{75} - P_{25})$

Límite de seguridad superior: $LSS = P_{75} + 1.5 \times (P_{75} - P_{25})$

El criterio para eliminar un dato atípico fue:

Si algún: $Xi < LSI$ o $Xi > LSS$, entonces se considera Xi un dato atípico y se elimina de la base de datos.

El resultado del preprocesamiento de datos fue la obtención de la base de datos consistente para aplicar los métodos de imputación propuestos en la investigación.

Paso 3. Prueba del mecanismo de los datos faltantes

Se aplicaron dos pruebas para verificar el mecanismo de los datos faltantes en los ingresos de los hogares, y la condición deseable era que los datos faltantes sigan un patrón MCAR o un MAR (Ver 2.1.3 la descripción de las pruebas). Para esto, se usaron las siguientes pruebas:

1. **Prueba univariada.** Se aplicaron pruebas t de la diferencia de medias independientes para todas las variables donantes. Suponiendo que los datos faltantes y completos provienen de la misma población, entonces se espera que las medias sean similares (Dixón, 1983). Para el caso de la imputación de la ENAHO, se usó la variable ingreso para definir los dos grupos: datos faltantes y datos no faltantes. Se aplica pruebas de "t" para la diferencia de medias independientes a todas las variables X's. Si la prueba resulta no significativa, hay evidencia estadística de que las medias son similares para los dos grupos (faltantes y no faltantes), lo cual indicaría que los datos faltantes tienen un mecanismo MCAR. Si resulta significativa la prueba, se sugiere que los datos faltantes siguen un MAR o MNAR (Ver 2.1.3, el procedimiento para aplicar las dos pruebas).

   Las hipótesis formuladas para la hipótesis nula y alternante son respectivamente:

   $H_0: \mu_{Obs} = \mu_{Mis}$

   $H_1: \mu_{Obs} \neq \mu_{Mis}$

   Dónde:

   μ_{Obs} Representa a la media de una variable donante (X), que corresponde al grupo de los datos que están completos en la variable Y.

μ_{Mis} Representa a la media de una variable donante (X), que corresponde al grupo de los datos faltantes en la variable Y.

2. **Prueba multivariada.** Se aplica la prueba propuesta por (Little, 1988). La prueba se basa en la prueba estadística de Máxima Verosimilitud. La prueba estadística es una suma ponderada de diferencias estandarizadas entre los grupos de medias y la media general, usando como estadístico de prueba una Chi-Cuadrado.

   Las hipótesis formuladas para la hipótesis nula y alternante son respectivamente:

   H_0: El mecanismo de los datos faltantes es un MCAR

   H_1: El mecanismo de los datos faltantes no es un MCAR

Paso 4. Aplicación de los métodos de imputación

Se propuso aplicar los siguientes métodos de imputación para reemplazar los datos faltantes en los ingresos de los hogares:

1) Método de eliminación de datos faltantes (listwise). Se analiza solo los datos completos, de tal manera que se elimina todos los registros de la base de datos que contengan datos faltantes en el ingreso.
2) Método de imputación por la media y la mediana. Se estima la media y mediana con los datos completos del ingreso y luego se reemplaza en los datos faltantes.
3) Método de imputación Hot-Deck aleatorio. Para cada dato faltante, se selecciona aleatoriamente un dato de los completos para ser reemplazado.
4) Método de imputación k vecino más cercano. Se selecciona de los datos completos los k vecinos más cercano para reemplazar a cada dato faltante mediante medir las distancias menores. Se propone usar como criterio de agregación la media y la mediana para reemplazar el dato faltante. La métrica de distancia que se usará es la Gower modificada (es la usada por el paquete VIM).

   Para aplicar este método se debe seleccionar el mejor valor de k. Para esto, con la base de datos completa se simula una muestra aleatoria de datos faltantes (missing) de tal manera que se tenga un mecanismo de datos faltantes MCAR. Luego, se aplica el método de imputación del k-vecino más cercano y se calcula los errores cuadrados medios entre los

observados e imputados y para valores de k entre 1 a 15. El valor de k corresponderá al menor error cuadrado medio hallado.

Paso 5. Comparación de los métodos de imputación

Con el propósito de comparar y evaluar los resultados con los métodos de imputación propuestos, se calculan una serie de estadísticas que permitan evaluar la performance y la precisión de los diferentes métodos de imputación, en base de los valores observados e imputados. Se determinan:

- Correlaciones entre los valores observados e imputados
- Diferencia de raíz cuadrado medio entre observados e imputados
- Cálculo de intervalos de confianza del 95% para la media

También es este caso, usando la base de datos completa se procedió a simular una muestra aleatoria de datos faltantes, de tal manera de conseguir que tengan un mecanismo MCAR. Con esta base de datos se aplicarán los métodos de imputación propuestos, y calculando los errores cuadrados medios y correlaciones entre los observados y los imputados.

Paso 6. Obtención de la base de datos completa. Se género y almacenó en un archivo la base de datos completa con los valores observados e imputados para los ingresos por hogar en la ENAHO 2017-tercer trimestre.

IV. RESULTADOS Y DISCUSIÓN

En la presente tesis se aplicó el método del k vecino más cercano para la imputación de datos faltantes de los ingresos de los hogares en la ENAHO 2017 tercer trimestre. Se comparan varios métodos para el tratamiento de datos faltantes; la eliminación de datos faltantes, los métodos de imputación usando la media y la mediana, el método Hot-Deck aleatorio y el k vecino más cercano usando como valor de agregación la media y la mediana. A continuación, se presenta los resultados del procedimiento estadístico propuesto en la metodología para satisfacer los objeticos definidos en esta investigación.

Diseño de la base de datos

Se realizó el enlace al Portal del INEI. Se recopiló la base de datos de la ENAHO 2017 – tercer trimestre que se encuentran en el formato .sav. En la Figura 1, se muestra el portal del INE y los módulos que se descargan: Módulo 100 (Información de la vivienda y el hogar), Módulo 300 (Información de la salud), Módulo 400 (Información de la educación) y Módulo 500 (Información del empleo e ingreso). Los cuatro archivos constituyen tablas relacionadas que se fusionaron para obtener la Base de Datos Integrada.

4.1 Preprocesamiento de datos

Con la finalidad de preparar los datos para aplicar el proceso de imputación de datos, se aplicaron técnicas del preprocesamiento de datos para la limpieza (manejos de datos faltantes y datos atípicos) y la transformación de los datos (creación de nuevas variables).

1) Definición de variables

Con el propósito de definir el conjunto de variables donantes (X) y la variable destino (Y), se generan nuevas variables a partir del conjunto de variables definidas en el punto 3.2.3. Así, se definieron las siguientes variables:

- **Variable destino.** Es la variable definida como el ingreso mensual total del hogar. Está variable se ha generado como la suma de los ingresos en los diferentes rubros: ocupación principal, ocupación secundaria, extraordinarios, transferencias corrientes del país, transferencias corrientes del extranjero, rentas de la propiedad y otros ingresos.

  Y = Ingresos mensual del hogar (soles).

- **Variables donantes.** Son las variables que sirvieron para llevar a cabo el proceso de imputación con el k vecino más cercano de los ingresos faltantes de los hogares. A partir de las variables propuestas se definieron 6 variables donantes. La variable X6, corresponde a la suma de los gastos que hacen las familias en alimentación y transporte.

  X1 = Total gasto mensual servicios de la vivienda (soles)

  X2 = Gasto por servicios de salud (soles)

  X3 = Número de horas trabajó a la semana en su ocupación principal

  X4 = Tiempo trabajando en la ocupación principal (años)

  X5 = Número de horas trabajadas a la semana en su ocupación secundaria

  X6 = Gasto en alimentación y transporte (soles)

2) Manejo de datos faltantes (missing)

Se analizaron las variables donantes X's con la finalidad de detectar datos faltantes. Se procedió a aplicar el procedimiento de eliminación de datos (listwise), que consiste en eliminar las observaciones (registros) que tuviera en alguna variable un dato faltante.

3) Manejo de datos atípicos (outliers)

Se realizó el tratamiento de datos atípicos para la variable destino Y, con la finalidad de mejorar el proceso de imputación. La base de datos inicialmente tiene 2324 hogares, siendo 1566 (67,4%) que tienen valores del ingreso y 758 (32,6%) presentan datos faltantes en el ingreso. Para la detección de los datos atípicos se aplicó el procedimiento del diagrama de cajas. En el Cuadro 6, se presenta medidas estadísticas para la variable Y del ingreso considerando solo los hogares con valores completos y los respectivos intervalos de seguridad inferior y superior para identificar los datos atípicos.

Cuadro 6. Medidas estadísticas para la variable ingreso

Medida estadística	**Y**
n	1566
Media	3543,5
Desviación estándar	25948,63
Mínimo	5,00
Máximo	10899,0
P_{25}	250,0
P_{50}	840,0
P_{75}	3055,0
ISI	-3957,5
ISS	7262,5

Siendo el valor del ISI igual a -3957,5, no se identifican datos atípicos pequeños; sin embargo, siendo el ISS igual a 7262,5 se identifican datos atípicos altos, al menos el máximo es igual a 10899,0. Por lo tanto, se eliminan 152 de las observaciones (registros) de la base de datos cuyo ingreso son mayores a 7262,5 que representa el 6,5% respecto al total de datos. Adicionalmente, se eliminaron registros con valores ceros en las variables X's.

Luego de aplicar el preprocesamiento se obtuvo una base de datos con un total de 1872 registros (hogares) para el proceso de imputación. En el Cuadro 7, se muestra la distribución del número de hogares con datos del ingreso faltante (458) y con datos completos (1414) que corresponden al 24,5% y 75,5% respectivamente.

Cuadro 7. Distribución de los hogares con ingresos faltantes y completos

Ingresos	**Número**	**Porcentaje**
Faltantes	458	24,5
Completos	1414	75,5
Total	1872	100,0

En el Cuadro 8, se muestran las medidas estadísticas para las seis variables donantes (X´s) para lo cual se considera el total de las observaciones (1872) y para la variable destino (Y) sólo los datos completos (1414).

Cuadro 8. Medidas estadísticas de las variables donantes (X's) y faltante (Y)

Variables	Media	Desviación estándar
X1	164,4	135,06
X2	101,9	342,63
X3	12,8	18,57
X4	41,1	19,51
X5	14,3	12,87
X6	50,1	13,34
Y	1431,5	1667,92

En el Cuadro 9, se muestran las medidas estadísticas para la Y del ingreso de los hogares, después de realizar todo el preprocesamiento de datos.

Cuadro 9. Medidas estadísticas del ingreso de los hogares

Medida estadística	Valores
n	1414
Media	1431,5
Mediana	640,0
Desviación estándar	1667,92
Mínimo	5,0
Máximo	7250,0
Rango intercuartil	1981,25
P25	220,0
P50	640,0
P75	2200,0
Asimetría	1,423

En la Figura 4, se muestra el histograma que representa la distribución de los ingresos de los hogares después de realizar el preprocesamiento de datos.

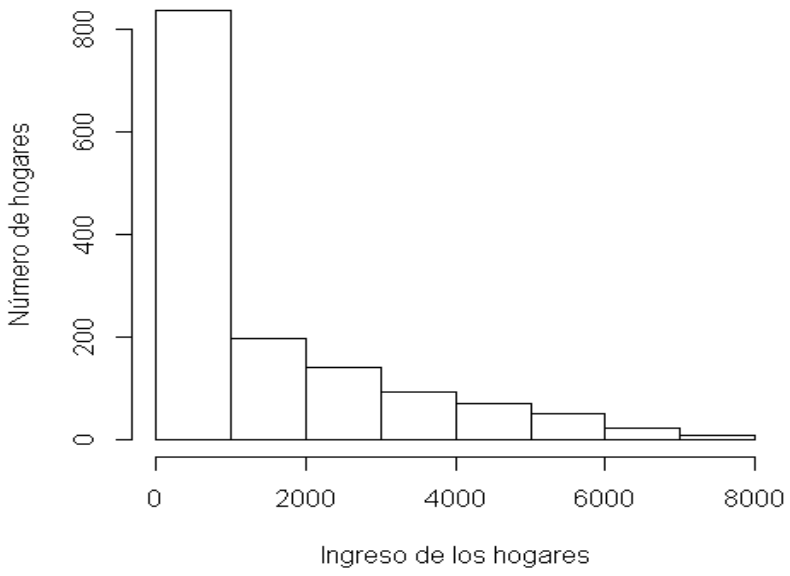

Figura 4. Distribución del ingreso de los hogares

4) Transformación de datos

Con la finalidad de aplicar los procedimientos de inferencia estadística, se aplicaron la transformación de datos a escala logaritmo en base 10 al conjunto de variables donantes y la variable destino. En la Figura 5, se presenta los histogramas y la curva normal con los datos

transformados con la función log base 10. Se puede observar que existe una aproximación a la Normal de las variables X's e Y. Esto justifica la aplicación del procedimiento de inferencia.

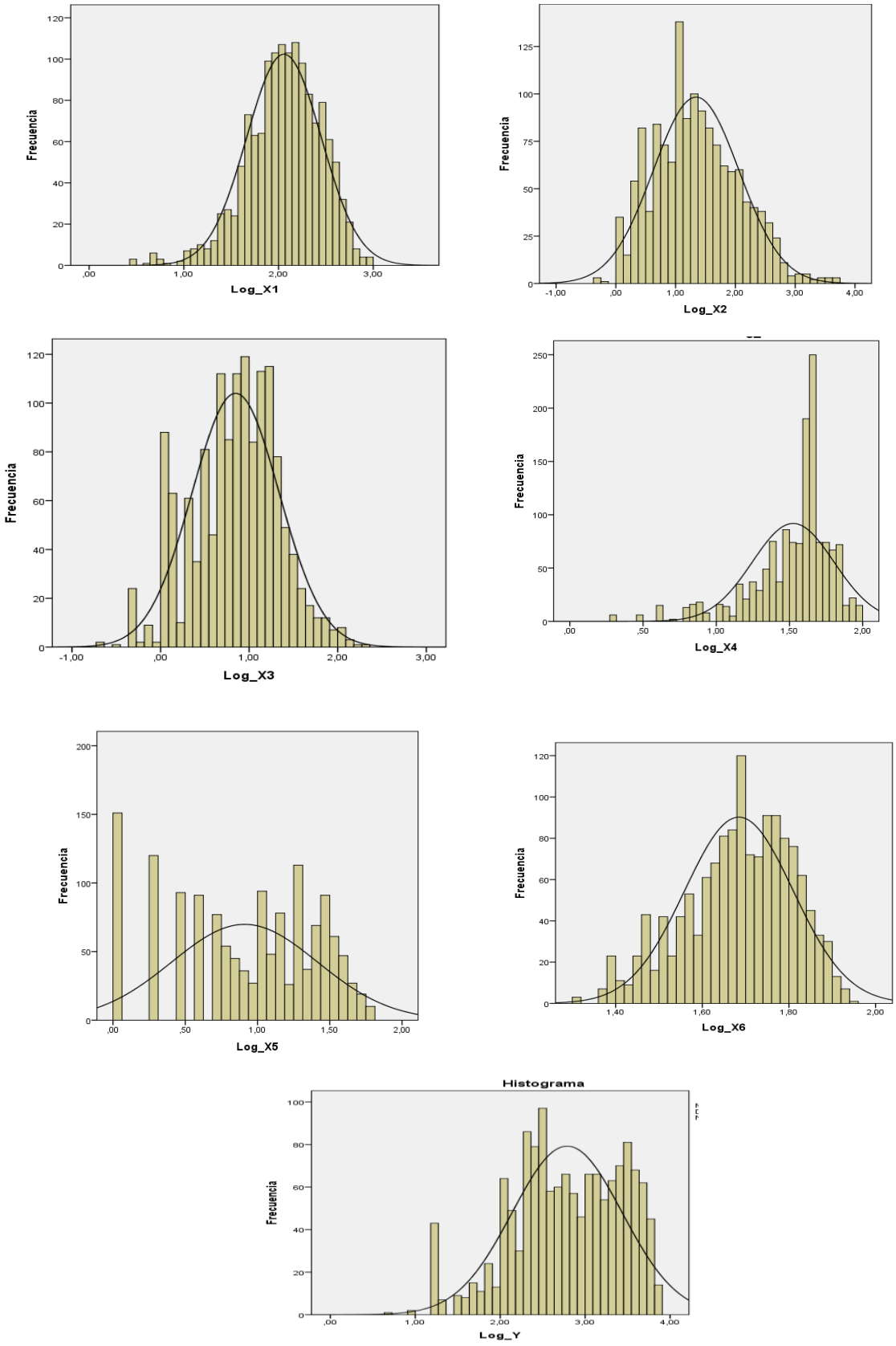

Figura 5. Histogramas de las variables con transformación Log10

4.2 Prueba del mecanismo de los datos faltantes

Existen una variedad de métodos y técnicas propuestas para el tratamiento de datos faltantes. Muchos de los métodos de imputación que se están utilizando para reemplazar los datos faltantes por estimaciones usando los datos existentes, suponen para su aplicación que el patrón de dichos datos faltantes sigue un comportamiento MCAR. En esta investigación se aplica el método univariado basado en la estadística "t" para la prueba de diferencia de medias independiente para cada variable donante (Díxon, 1983) y la prueba multivariada basada en la estimación de máxima verosilimitud (Little, 1988).

La prueba univariada plantea como hipótesis nula, que los datos faltantes siguen un patrón MCAR frente a la alterna que correspondería a un patrón MAR o NMAR. Para la aplicación de la prueba "t" para la diferencia de medias independientes, se considera según la variable destino (Y) que los dos grupos a ser evaluados son: datos faltantes y no faltantes. En el Cuadro 10, se muestra las medias y las desviaciones estándar para cada variable donantes agrupadas según el ingreso con datos faltantes y no faltantes. Se observa que algunas variables muestran una diferencia grande entre la media del grupo de datos faltantes y los no faltantes. Esto será corroborado cuando se presente los resultados de las pruebas "t" para la diferencia de medias independientes.

Cuadro 10. Medias estadísticas para las variables donantes agrupadas por datos faltantes y no faltantes

Variables	Grupo	n	Media	Desviación estándar
X1	Faltante	458	170,7	128,57
	No faltante	1414	162,4	137,08
X2	Faltante	458	102,9	320,14
	No faltante	1414	101,6	349,71
X3	Faltante	458	13,1	21,72
	No faltante	1414	12,8	17,44
X4	Faltante	458	46,8	21,21
	No faltante	1414	39,2	18,56
X5	Faltante	458	14,9	11,83
	No faltante	1414	14,1	13,19
X6	Faltante	458	49,4	12,15
	No faltante	1414	50,3	13,69

En el Cuadro 11, se presentan los resultados de la prueba t para diferencia de medias con las respectivas pruebas calculadas y los p-valores. Previamente los resultados de la prueba F de

Levene, resultaron con variancias homogéneas para las variables X1 y X3 y heterogéneas X2, X4, X5 y X6. Respecto a la prueba t para diferencias de medias independientes, se puede afirmar con un nivel de significación del 5%, que sólo la variable X4 resultó significativa. Mientras se puede concluir que las variables X1, X2, X3, X5 y X6 no hay evidencia estadística para rechazar que las medias en las variables donantes entre los datos faltantes y no faltantes son similares. Este resultado indica que existe un mecanismo MCAR en los datos faltantes del ingreso de los hogares.

Cuadro 11. Prueba "t" diferencia de medias para probar el mecanismo de los datos faltantes

Variables	t	p-valor
X1	2,50	0,013
X2	1,73	0,083
X3	0,63	0,546
X4	5,43	0,000
X5	3,27	0,001
X6	-0,85	0,395

En el Cuadro 12, se presenta los resultados de la prueba multivariada para evaluar el mecanismo de los datos faltantes. La prueba global indica con un nivel de significación del 5%, que hay evidencia estadística para rechazar que existe un patrón MCAR en los datos faltantes; sin embargo, las pruebas individuales indican que los datos faltantes siguen un patrón MCAR

Cuadro 12. Prueba multivariada para evaluar el patrón de los datos faltantes

Variables	Chi-Cuadrado	p-valor
X1	5,68	0,017
X2	2,99	0,083
X3	0,36	0,546
X4	9,08	0,023
X5	10,61	0,001
X6	0,66	0,417
Global	50,93	0.000

Por lo tanto, los resultados de la prueba individual y multivariada indican que el mecanismo de los datos faltantes en los ingresos de los hogares es un MCAR. Este resultado, permite que los métodos de imputación propuestos en este trabajo sean aplicados con la justificación que

ellos requieren de un mecanismo MCAR; es decir, que la ocurrencia de los datos faltantes en los ingresos de los hogares no depende de las variables X's ni de la misma variable ingreso.

4.3 Aplicación de los métodos de imputación

En el presente trabajo de investigación se aplicó y se comparó el método de eliminación de datos faltantes y el método de imputación por la media y la mediana, el método de imputación Hot-Deck y el método de imputación por el k vecino más cercano. A continuación, se presenta los respectivos análisis y los resultados encontrados para cada uno de los métodos propuestos.

4.3.1 Método por eliminación

El método más práctico y sencillo, es la eliminación por lista (ListWise) y su aplicación se basa en el supuesto de que los datos faltantes corresponden a un MCAR. El procedimiento consiste en eliminar todas las observaciones que presenta valores faltantes en alguna de las variables, obteniendo un conjunto de datos con valores completos en todas las variables. En el Cuadro 13, se presenta las medidas estadísticas después de eliminar las observaciones faltantes (458), quedando 1414 observaciones con datos completos.

Cuadro 13. Medidas estadísticas (Método eliminación)

Variables	Media	Desviación estándar
X1	162,4	137,08
X2	101,6	349,71
X3	12,8	17,44
X4	39,2	18,56
X5	14,1	13,19
X6	50,3	13,69
Y	1431,5	1667,92

En el Cuadro 14, se presenta las medidas estadísticas calculadas para el ingreso de los hogares usando el método de eliminación.

Cuadro 14. Medidas estadísticas del ingreso de los hogares (Método eliminación)

Medida estadística	Valores
Media	1431,5
Mediana	640,0
Desviación estándar	1667,92
Rango intercuartil	1981,25

P25	220,0
P50	640,0
P75	2200,0
Asimetría	1,423

En la Figura 5, se muestra el histograma y la forma de distribución que presenta los ingresos de los hogares usando el procedimiento de eliminación de observaciones. Como se aprecia, se evidencia una distribución asimétrica a la derecha (asimetría positiva).

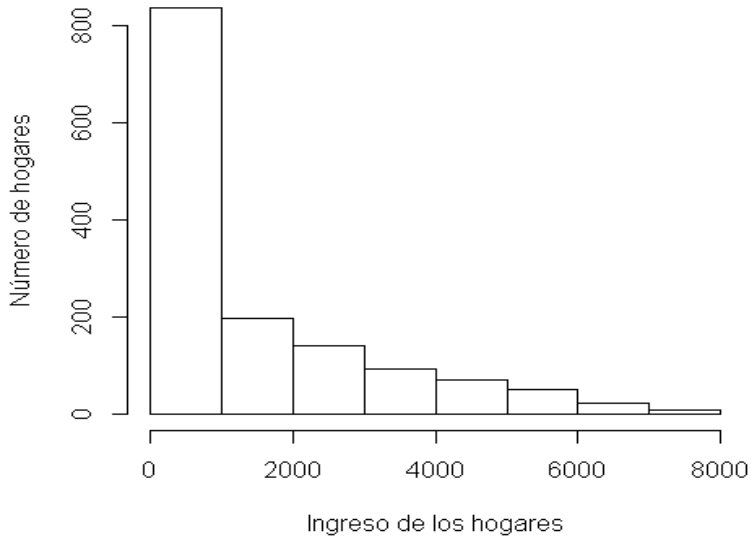

Figura 6. Distribución del ingreso de los hogares (Método eliminación)

En el Cuadro 15, se presenta los intervalos de confianza del 95% para la media y la desviación estándar de los ingresos de los hogares, luego de aplicar el método de la eliminación de datos.

Cuadro 15. IC del 95% para la media y la desviación estándar del ingreso (Método eliminación)

Medida estadística	Límite inferior	Límite superior	Amplitud
Media	1344,47	1518,49	174,02
Desviación estándar	1608,63	1731,78	123,14

En el Cuadro 16, se presenta los coeficientes de correlación de Pearson entre la variable Y y las X´s con sus respectivos niveles de significación. Se aprecia que todas las variables muestran una correlación significativa con la variable ingreso.

Cuadro 16. Correlaciones entre Y y las X´s (Método de eliminación)

	X1	X2	X3	X4	X5	X6
Y	0,455	0,062	0,268	0,204	-0,068	-0,132
p-Valor	0.000	0.020	0.000	0.000	0,011	0,000

4.3.2 Método de la imputación por la media y mediana

Se realiza la imputación por el método más simple que es el reemplazar los datos faltantes por la media o mediana de los datos completos. Con los datos completos del ingreso, se calculó que su promedio es igual a 1431,5. Este valor de la media es usado para reemplazar en todos los datos faltantes de los ingresos de los hogares. En el Cuadro 17, se presenta las medidas estadísticas resultantes usando la imputación por la media.

Cuadro 17. Medidas estadísticas del ingreso de los hogares (Método de imputación por la media)

Medida estadística	**Valores**
Media	1431,5
Mediana	1431,5
Desviación estándar	1449,47
Rango intercuartil	1200,00
P_{25}	300,00
P_{50}	1431,5
P_{75}	1500,00
Asimetría	0

En la Figura 6, se muestra el histograma y la forma de distribución que presenta los ingresos de los hogares usando el método de imputación de la media. Como se aprecia, se evidencia una distribución asimétrica a la derecha (asimetría positiva).

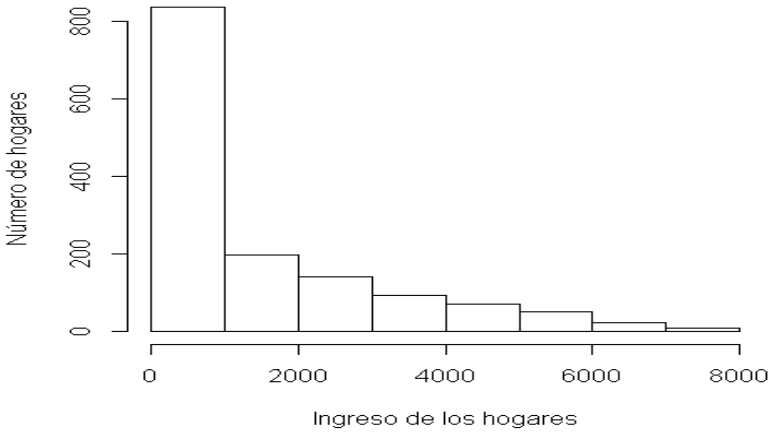

Figura 7. Distribución del ingreso de los hogares (Método de imputación de la media)

En el Cuadro 18, se presenta los intervalos de confianza del 95% para la media y la desviación estándar de los ingresos de los hogares, luego de aplicar el método de imputación por la media.

Cuadro 18. IC del 95% para la media y la desviación estándar del ingreso (Método de imputación por la media)

Medida estadística	Límite inferior	Límite superior	Amplitud
Media	1365,78	1497,19	131,41
Desviación estándar	1404,49	1497,46	92,97

En el Cuadro 19, se presentan los coeficientes de correlación y los respectivos p-valor entre las variables X's y los ingresos imputados.

Cuadro 19. Correlaciones entre Y y las X´s (Método de imputación por la media)

	X1	X2	X3	X4	X5	X6
Y	0,401	0,055	0,219	0,169	-0,060	-0,118
p-Valor	0.000	0.020	0.000	0.000	0,011	0,000

Con los datos completos del ingreso, se calculó que la mediana es igual a 640,0. Este valor de la mediana es usado para reemplazar en todos los datos faltantes de los ingresos de los hogares. En el Cuadro 20, se presenta las medidas estadísticas resultantes usando la imputación por la mediana.

Cuadro 20. Medidas estadísticas del ingreso de los hogares (Método de imputación por la mediana)

Medida estadística	Valores
Media	1237,84
Mediana	640,00
Desviación estándar	1488,89
Rango intercuartil	1200,00
P25	300,00
P50	640,00
P75	1500,00
Asimetría	1,204

En la Figura 7, se muestra el histograma y la forma de distribución que presenta los ingresos de los hogares usando el método de imputación de la mediana. Como se aprecia, se evidencia una distribución asimétrica a la derecha (asimetría positiva).

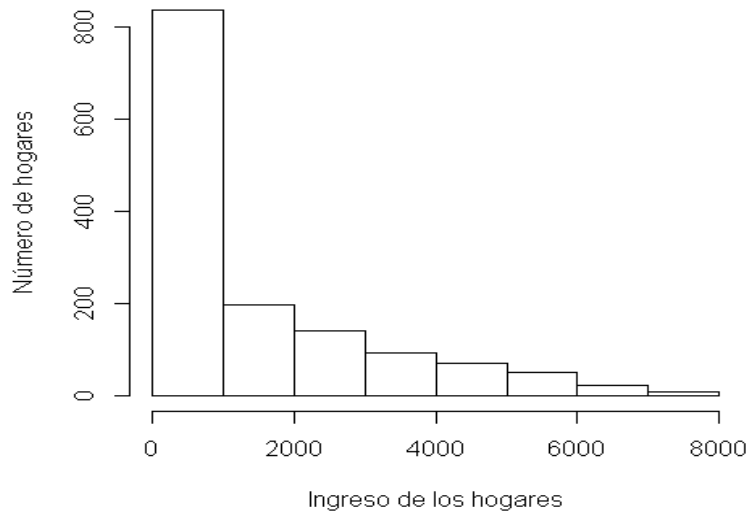

Figura 8. Distribución del ingreso de los hogares (Método de imputación de la mediana)

En el Cuadro 21, se presenta los intervalos de confianza del 95% para la media y la desviación estándar de los ingresos de los hogares, luego de aplicar el método de imputación por la mediana.

Cuadro 21. IC del 95% para la media y la desviación estándar del ingreso (Método de imputación por la mediana)

Medida estadística	Límite inferior	Límite superior	Amplitud
Media	1170,35	1305,33	134,98
Desviación estándar	1442,68	1538,18	95,50

En el Cuadro 22, se presentan los coeficientes de correlación y los respectivos p-valor entre las variables X's y los ingresos imputados.

Cuadro 22. Correlaciones entre Y y las X´s (Método de imputación por la mediana)

	X1	X2	X3	X4	X5	X6
Y	0,384	0,053	0,211	0,126	-0,065	-0,108
p-Valor	0.000	0.022	0.000	0.000	0,005	0,000

4.3.3 Método de imputación Hot-Deck

Se aplica la imputación por el método Hot-Deck con reemplazo aleatorio para reemplazar los datos faltantes del ingreso de los hogares. En el Cuadro 23, se presenta las medidas estadísticas para el ingreso de los hogares usando la imputación Hot-Deck.

Cuadro 23. Medidas estadísticas del ingreso de los hogares (Método de imputación Hot-Deck)

Medida estadística	Valores
Media	1082,0
Mediana	300,00
Desviación estándar	1574,59
Rango intercuartil	1484,00
P_{25}	16,00
P_{50}	300,00
P_{75}	1500,00
Asimetría	1,489

En la Figura 8, se muestra el histograma y la forma de distribución que presenta los ingresos de los hogares usando el método de imputación Hot-Deck aleatorio. Como se aprecia, se evidencia una marcada distribución asimétrica a la derecha (asimetría positiva) del ingreso.

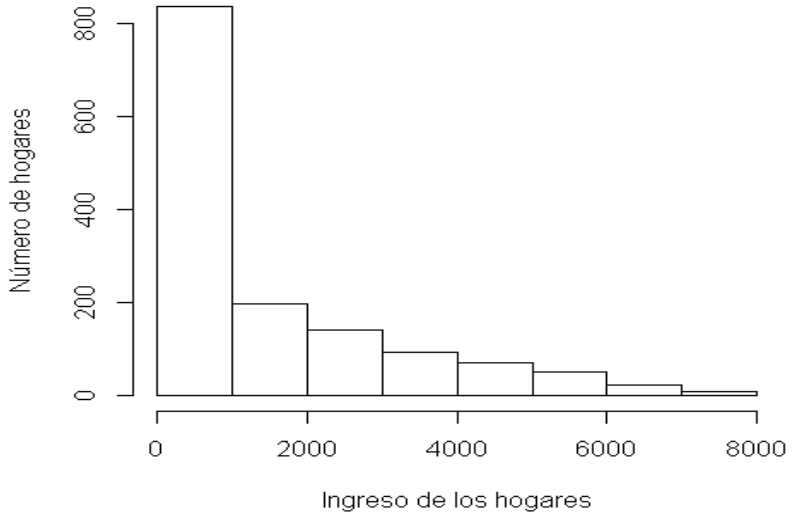

Figura 9. Distribución del ingreso de los hogares (Método de imputación Hot-Deck aleatorio)

En el Cuadro 24, se presenta los intervalos de confianza del 95% para la media y la desviación estándar de los ingresos de los hogares, luego de aplicar el método de imputación Hot-Deck.

Cuadro 24. IC del 95% para la media y la desviación estándar del ingreso (Método de imputación Hot-Deck)

Medida estadística	Límite inferior	Límite superior	Amplitud
Media	1010,13	1152,88	142,75
Desviación estándar	1525,72	1626,71	100,99

En el Cuadro 25, se presentan los coeficientes de correlación y los respectivos p-valor entre las variables X's y los ingresos imputados.

Cuadro 25. Correlaciones entre Y y las X´s (Método de imputación Hot-Deck)

	X1	X2	X3	X4	X5	X6
Y	0,359	0,050	0,199	0,089	-0,066	-0,097
p-Valor	0.000	0.020	0.000	0.000	0,011	0,000

4.3.4 Método de imputación k vecino más cercano

Se aplica la técnica del k vecino más cercano como método de imputación para reemplazar los datos faltantes de los ingresos de los hogares. Antes de aplicar el método de imputación, se selecciona el conjunto de variables donantes del conjunto de las seis variables con datos completos (X´s) y luego se seleccionó el mejor valor de k.

1) Selección de las variables donantes

En el Cuadro 26, se presentan los coeficientes de correlación y el p-valor de las variables con los datos completos (X's) con la variable ingreso que contiene datos faltantes (Y). Considerando un nivel de significación del 1%, se observa que las variables que muestran una correlación significativa con la Y son X1, X3, X4, X6 y por lo tanto constituirán el conjunto de variables donantes.

Cuadro 26. Correlaciones entre Y y las X´s para seleccionar las variables donantes

	X1	X2	X3	X4	X5	X6
Y	0,455	0,062	0,268	0,204	-0,068	-0,132
p-Valor	0.000	0.020	0.000	0.000	0,011	0,000

2) Selección del valor k

Para seleccionar el mejor valor k, se generó en primer lugar con el conjunto de datos completos una muestra aleatoria simulada del 20% con los datos faltantes (80% no faltante) a fin de tener un mecanismo MCAR. En el Cuadro 27, se presenta la distribución de la muestra simulada de los datos faltantes y no faltantes a partir de los datos completos.

Cuadro 27. Distribución de la muestra simulada

Muestra	Número	Porcentaje
Faltantes	267	18,9
No faltantes	1147	81,1
Total	1414	100,0

El análisis de sensibilidad para diferentes valores de k es evaluado a partir del Error Cuadrado Medio (ECM). En el Cuadro 28, se presenta para los valores de los ECM calculados para k entre 1 a 1. Por consiguiente, el valor de k=9 es el seleccionado puesto que resultó con el menor ECM (1334,2).

Cuadro 28. Análisis de sensibilidad para diferentes valores k

Valores de k	ECM
1	1867,0
2	1580,8
3	1533,8
4	1506,6
5	1447,9
6	1423,7
7	1385,3
8	1356,4
9	**1334,2**
10	1346,5
11	1342,6
12	1343,4
13	1341,9
14	1344,3
15	1345,9

En la Figura 9, se muestra el resultado del análisis de sensibilidad simulando una muestra aleatoria de datos faltantes considerando los datos completos. Se observa que hay decaimiento

fuerte del ECM cuando para los ocho primeros valores de k, llegando al menor valor para k=9 y permaneciendo a partir de este valor casi constante. Por lo tanto, este resultado permite seleccionar como mejor valor k=9.

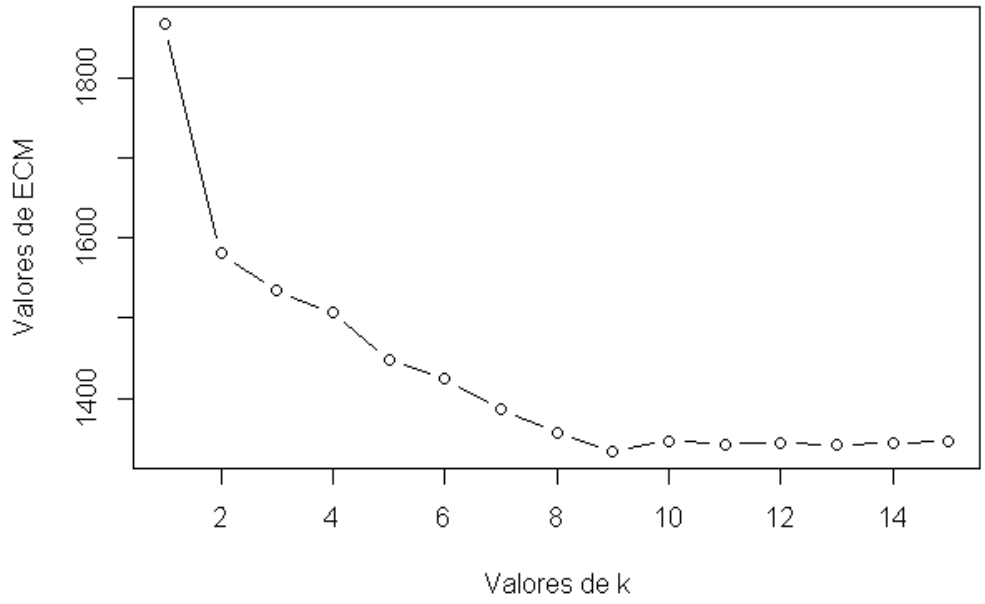

Figura 10. Análisis de sensibilidad valorando el ECM para diferentes valores de k

Se aplicó la imputación k vecino más cercano, con el valor que resultó siendo el mejor k=9, con variables donantes X1, X3, X4, X6, con métrica de distancia de Gower modificada y usando la media y mediana como medidas de agregación. En el Cuadro 29, se presenta las medidas estadísticas para el ingreso de los hogares usando la imputación k vecino más cercano usando la media y la mediana.

Cuadro 29. Medidas estadísticas del ingreso de los hogares (Método de imputación k vecino más cercano con la media y la mediana)

Medida estadística	**Media**	**Mediana**
Media	1425,0	1335,2
Mediana	857,2	644,5
Desviación estándar	1510,9	1537,9
Rango intercuartil	1796,9	1693,0
P_{25}	280,0	250,0
P_{50}	857,2	644,5
P_{75}	2076,9	1943,0
Asimetría	1,127	1,347

En la Figura 10, se muestra el histograma y la forma de distribución que presenta los ingresos de los hogares usando el método de imputación k vecino más cercano. Como se aprecia, se evidencia una marcada distribución asimétrica a la derecha (asimetría positiva) del ingreso.

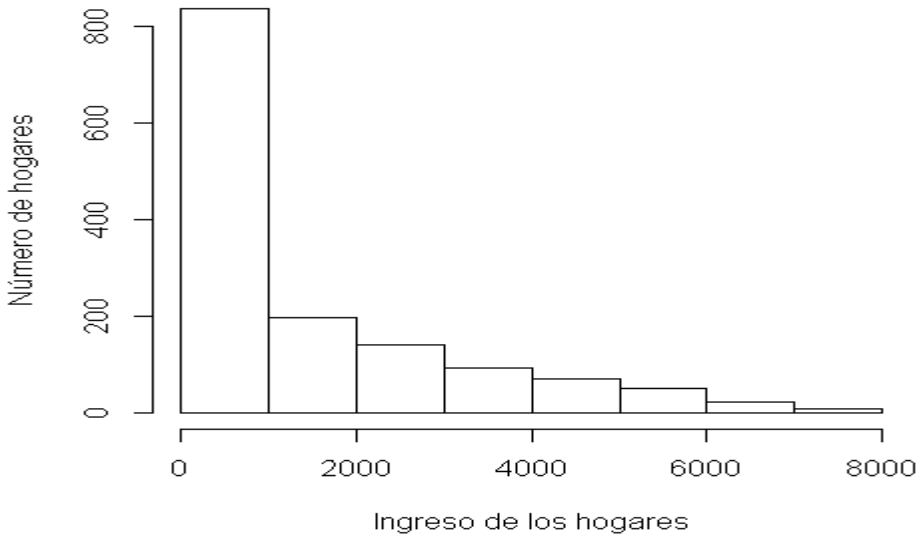

Figura 11. Distribución del ingreso de los hogares (Método de imputación k vecino más cercano)

En el Cuadro 30, se presenta los intervalos de confianza del 95% para la media del ingreso de los hogares considerando el método de imputación del k vecino más cercano. Se puede apreciar que con el método de imputación k vecino más cercano las estimaciones por intervalos de confianza para la media del ingreso de los hogares tienen amplitudes similares.

Cuadro 30. IC del 95% para la media del Ingreso de los hogares (Método de imputación k vecino más cercano)

Medida	LCI	LCS	Amplitud
Media	1356,5	1493,5	136,9
Mediana	1265,5	1404,9	139,4

En el Cuadro 31, se presenta los intervalos de confianza del 95% para la desviación estándar del ingreso de los hogares considerando el método de imputación del k vecino más cercano. Se puede apreciar que con el método de imputación k vecino más cercano las estimaciones por intervalos de confianza para la desviación estándar del ingreso de los hogares tienen amplitudes similares.

Cuadro 31. IC del 95% para la desviación estándar del Ingreso de los hogares (Método de imputación k vecino más cercano)

Medida	LCI	LCS	Amplitud
Media	1464,1	1560,9	96,9
Mediana	1490,1	1588,8	98,6

En el Cuadro 32, se presentan los coeficientes de correlación y los respectivos p-valor entre las variables X's y los ingresos imputados.

Cuadro 32. Correlaciones entre Y y las X´s (Método de imputación k vecino más cercano)

		X1	X3	X4	X6
Media	Y	0,479	0,248	0,205	-0,139
	p-Valor	0.000	0.000	0.000	0,000
Mediana		0,483	0,253	0,165	-0,122
	p-Valor	0.000	0.000	0.000	0,000

4.4 Evaluación y comparación de los métodos de imputación

Para realizar la comparación y evaluación de los métodos de imputación propuestos en esta investigación, se consideró como la base de datos para el análisis el conjunto de datos completos. A partir de esta base de datos se simuló una muestra aleatoria del 10% con los datos faltantes (90% no faltante) a fin de tener un mecanismo MCAR. En el Cuadro 33, se presenta la distribución de la muestra simulada de los datos faltantes y no faltantes a partir de los datos completos.

Cuadro 33. Distribución de la muestra simulada

Muestra	Número	Porcentaje
Faltantes	126	8,9
No faltantes	1288	91,1
Total	1414	100,0

En el Cuadro 34, se presenta los resultados de los cinco métodos de imputación desarrollados en esta investigación. Se consideran como medidas para la comparación los ECM y los respectivos coeficientes de correlación entre los valores observados e imputados para cada uno de los métodos de imputación. Se observa que el método de imputación k vecino más cercano es el presenta los menores ECM. Siendo mucho menor el ECM cuando se usa la

mediana (444,4) que cuando la media (1412,6). Esta evaluación permite corroborar la suposición que el método de imputación con k vecino más cercano permite obtener una mayor precisión en las estimaciones. Respecto a los coeficientes de correlación entre los valores observados y los imputados, el método de imputación del k vecino más cercano tuvo una pequeña diferencia mayor en comparación de los otros métodos.

Cuadro 34. Comparación de los métodos de imputación con los ECM y correlaciones

Método de imputación	ECM	Correlaciones
Por la media	1504,5	0.963
Por la mediana	1619,9	0,958
Hot-Deck aleatorio	1963,7	0,940
k vecino más cercano con la media	1412,6	0,968
k vecino más cercano con la mediana	444,4	0,964

En la Figura 11, se muestra la comparación de los ECM de los cinco métodos de imputación analizados. Se aprecia que el método k vecino más cercano presenta los menores ECM, aunque el uso de la mediana resulta con mucho menor ECM.

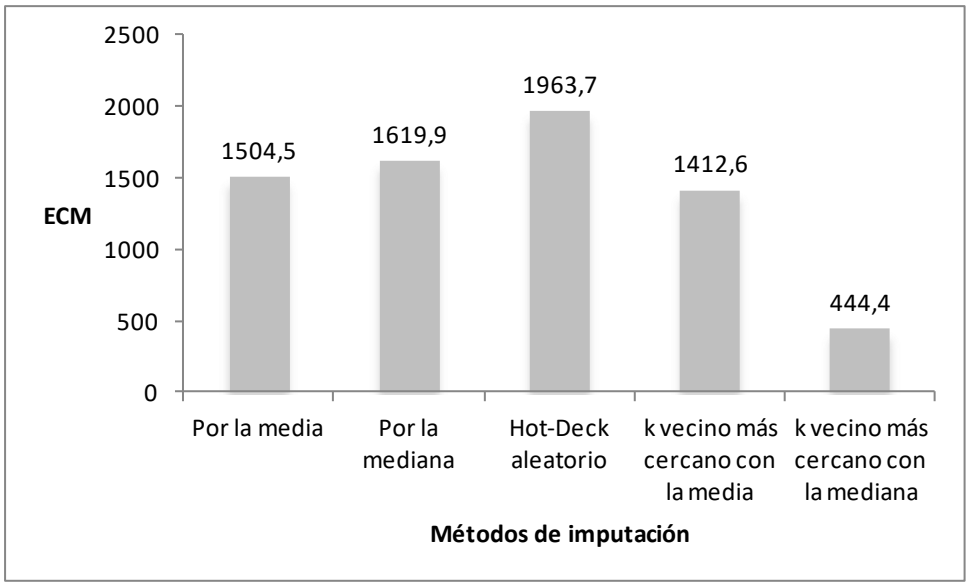

Figura 12. Comparación de los valores ECM

En el Cuadro 35, se presenta el resumen para comparar los intervalos de confianza del 95% para estimar el ingreso medio de los hogares.

Cuadro 35. Comparación de IC del 95% para la media del ingreso de los hogares

Método de imputación	LI	LS	Amplitud
Eliminación de datos	1344,47	1518,49	174,02
Por la media	1365,78	1497,19	131,41
Por la mediana	1170,35	1305,33	134,98
Hot-Deck aleatorio	1356,5	1493,5	136,9
k vecino más cercano con la media	1265,5	1404,9	136,9
k vecino más cercano con la mediana	1265,5	1404,9	139,4

Cuadro 36. Comparación de IC del 95% para la desviación estándar del ingreso de los hogares

Método de imputación	LI	LS	Amplitud
Eliminación de datos	1608,63	1731,78	123,14
Por la media	1404,49	1497,46	92,97
Por la mediana	1442,68	1538,18	95,50
Hot-Deck aleatorio	1525,72	1626,71	100,99
k vecino más cercano con la media	1525,72	1626,71	100,99
k vecino más cercano con la mediana	1490,1	1588,8	98,6

V. CONCLUSIONES

Las conclusiones de la presente investigación son:

1. El método de imputación por k vecino más cercano, permitió identificar de las seis variables propuestas como donantes a cuatro. Se determinó con la generación de una muestra aleatoria con un 20% con datos faltantes artificiales y calculando el ECM para valores de k entre 1 a 15, que el mejor valor de k es 9 con el menor ECM de 1412,6.

2. Los datos faltantes en los ingresos mensuales de los hogares en la ENAHO 2017 trimestre 3, mostraron un mecanismo de MCAR (Missing Completing Aleatore Random). La prueba univariada y multivariada evidenciaron con un nivel de significación del 5% que la ocurrencia de los datos faltantes en los ingresos de los hogares no está relacionada o no depende de las variables donantes ni del mismo ingreso.

3. La estimación por intervalo de confianza del 95% para la media del ingreso mensual de los hogares para la ENAHO 2017 al aplicar los métodos de imputación, resultaron con similares amplitudes, siendo el menor obtenido con la media S/. 131,41; mientras el método del k vecino más cercano por la media y la mediana fueron iguales a S/.136,9 y S/.139,4 respectivamente. Las amplitudes de los intervalos de confianza del 95% para desviación estándar de los ingresos imputados fueron un poco diferentes, el menor valor fue con la media S/.92,97; mientras que por el método del k vecino más cercano por la media y la mediana fueron de S/.100,99 y S/.98,6 respectivamente.

4. El método de imputación por k vecino más cercano mostró una mejor precisión en la estimación de los valores imputados para los ingresos mensuales de los hogares. El método del k vecino más cercano obtuvo los menores valores del ECM, para con la media 1412,6 y con la mediana 444,4. Los coeficientes de correlaciones resultaron con valores muy similares, para k vecino más cercano 0,968 con la media y 0,964 con la mediana.

VI. RECOMENDACIONES

1. Considerar en el proceso de la imputación usar variables de agrupamiento que permitan tener mayor homogeneidad en los datos, de tal manera que aplique el proceso de imputación dentro de cada grupo. En el caso de la ENAHO la variable Dominio (Costa, Sierra, Selva, Lima, etc.) permite definir las clases o grupos con mayos homogeneidad en cuanto a las variables sociales, económicas, etc.

2. Usar la imputación múltiple como un método más avanzado para solucionar el problema de la no respuesta en las encuestas. La imputación múltiple, está demostrando mejorar la precisión de las estimaciones en muchos casos.

VII. REFERENCIA BIBLIOGRÁFICA

Batista, G. E., & Monard, M. (2002). A Study of K-Nearest Neighbour as a Imputation Method. pp. 251-260.

Batista, G., & Monard, M. C. (2002). An Analysis of Four Missing Data Treatment Methods for Supervised Learning.

Chen, J., & Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. Journal of Official Statistics, vol. 16, No. 2., pp. 113-131.

Cochran, W. G. (1977). Sampling Techniques, Second Edition. John Wiley and Sons, Inc.

De Leeuw, E. D. (2001). Reducing Missing Data in Surveys: An Overview of Methods. Quality and Quantity, vol. 35., pp. 147-160.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). Journal of Royal Statistical Society, B39, pp. 1-38.

Dixon, W. J. (1983). BMDP Statistical Software. Bereley:University of California Press.

Donza, E. (2013). Método de imputación de la no respuesta en las preguntas de ingresos en la Encuesta Permanente de Hogares. Gran Buenos Aires 1990-2010. Buenos Aires.

Downey, R. G., & King, C. V. (1998). Missing Data in Likert Ratings: A Comparison of Replacement Mathods. Journal of General Psychologu, pp. 175-191.

Eskelson, B., Tenmesgen, H., Lemay, V., Barret, T., & Crookston, N. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. USDA Forest Service, pp. 235-246.

García-Laencina, P., Sancho-Gómez, J., Figueiras-Vidal, A., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputting72., pp. 1483-1493.

Giocoechea, Aitor (2002). “Imputación basada en árboles de clasificación”. Eustat.

INEI (2015). Recuperado: https://www.cepal.org/sites/default/files/events/files/2015-10-tallereh-d-lucia-gaslac.pdf

Jonsson, P., & Wohlin, C. (2006). Benchmarking k-Nearest Neighbour Imputation with Homogeneous Likert Data. Empirical Sotgware Engineering: An International Journal, pp. 463-489.

LeMay, V., & Temesgen, H. (2004). Comparison of Nearest Neighbor Methods for Estimating Basal per Hectare Using Aerial Auxiliary Variables. Society of American Forestrs, pp. 209-219.

LeMay, V., & Temesgen, H. (2005a). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science 51, pp. 109-199.

Lindenboim, J., Graña, J., & Kennedy, D. (2006). Concepto, medición y utilidad de la distribución funcional del ingreso. Argentina 1993–2005.

Little, J. A. (1988). A Test of Missing Completely at Random for Multivariate Data With Missing Values . Journal of the American Statistical Association, Vo. 83, No. 404, pp. 1198-1202.

Little, R., & Rubin, D.B. (2002). Statistical Analysis with Missing Data. Wiley-Interscience.

Little, T.D., Jorgensen, T., Lang, K., & Moore, E. (2014). On the Joys of Missing Data. Journal of Pediatric Psychology, 39(2), pp. 151-162.

Lohr, S. (1999). Muestro: Diseño y Análisis. Thomson.

Mc Roberts, R. E., Nelson, M. D., & Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. pp. 457-468.

McRoberts, R. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment, pp. 489-499.

Medina, F., & Galván, M. (2007). Imputación de datos: teoría y práctica. Chile: Publicación de las Naciones Unidas.

Mesa Ávila, Dulce Maria y Useche Castro, Lelly María (2006). Una introducción a la Imputación de Valores Perdidos Terra Nueva Etapa, vol. XXII, núm. 31, pp. 127-151

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London, Series A.

Neyman, J., & Pearson, E.S. (1933). On the problem of most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A.

Packale, P., & Maltamo, M. (2007). The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment, pp. 328-341.

Polo, C., Behar, R., & Olaya, J. (2000). Comparación Empírica de la Eficiencia de Algunas Técnicas de Tratamiento de datos Faltantes Aplicadas al Análisis de Regresión Lineal Múltiple.

Raaijmakers, Q. A. (1999). ffectiveness of Different Missing Data Treatments in Surveys with Likert-Type Data. Educational and Psychological Measurement, vol. 59, No. 5., pp. 725-748.

Restrepo Estrada, M. I., & Marín Diazaraque, J. M. (2012). Imputación de ingresos en la Gran Encuesta Integrada de Hogares (geih) de 2010. Desarrollo y Sociedad, 219-243.

Rubin, D. B. (1976). Inference and missing data. Biometrika, Vol.63, pp. 581-592.

Salvia, A., & Donza, E. (1999). Salvia, A. y Donza, E. (1999). “Problemas de medición y sesgos de estimación derivados de no respuesta a las preguntas de ingresos en la Encuesta Permanente de Hogares (1990-1998)”. Revista de la Asociación Argentina de Especialistas en Estudios del Traba. Revista de la Asociación Argentina de Especialistas en Estudios del Trabajo, n.° 18, Buenos Aires.

Sande, I. (1982). Imputation in Surveys: Coping with reality. The American Statistician, Vol.36, 3, 145-152.

Schafer, J. (1990). Analysis of Incomplete Multivariate Data. Series Monographs on Statistics and Applied Probability. Chapman & Hall, Londres.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of art. Psychological Methods. Vol. 7, N°2., pp. 17-147.

Schafer, J., & Schenker, N. (1997). Inference with Imputed Conditional Means.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. Biometrics & Biostatistics, pp. 1-6.

Schmitt, P., Mendel, J., & Guedy, M. (2015). A Comparison of Six Methods for Missing Data Imputation. Biometrics & Biostatistics, vol.6, pp. 1-6.

Song, Q., Shepperd, M., & Cartwright, M. (2005). A Short Note on Safest Default Missingness Mechanism Assumptions. Empirical Software Engineering, Vol. 10, pp. 235-243.

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2016). VIM: Visualization and Imputation of. URL https://CRAN.R-project.org/package=VIM.

Todeschini, R. (1990). Weighted k-nearest neighbour method for the calculation of missing values. Chenometrics and Intelligent Laboratory Systems 9.201-205.

Tuominen, S., Fish, S., & Poso, S. (2003). Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. pp. 624-634.

Tutz, G., & Ramzan, S. (2014). Improved Methods for the Imputation of Missing Data by Nearest Neighbor Methods.

Zainuri, N., Jemain, A., & Mura, N. (2015). A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. pp. 449-456.

VIII. ANEXOS

Anexo 1. Resultados del programa con R para la imputación de datos: Package VIM

```
#  Programa con R para la imputación de datos. Package VIM
#  install.packages("VIM")
> library("VIM")
> #  Lectura de archivo de datos
> Datos=read.table("Collazos_Enaho_2017_03_Final.csv",header=TRUE,sep=",")
> attach(Datos)
The following objects are masked from Datos (pos = 3):

    X1, X2, X3, X4, X5, X6, Y

The following objects are masked from Datos (pos = 4):

    X1, X2, X3, X4, X5, X6, Y

The following objects are masked from Datos (pos = 5):

    X1, X2, X3, X4, X5, X6, Y

> names(Datos)# Ver los nombres de las variables
[1] "X1" "X2" "X3" "X4" "X5" "X6" "Y"
> str(Datos)# Ver número de observaciones, variables y algunos datos
'data.frame':   1872 obs. of  8 variables:
 $ X1: num  212 461 80 139 405 227 133 225 168 164 ...
$ X2: num  240 18 120 6 600 40 5 170 248 13 ...
 $ X3: num  19.5 12.7 17.5 10 2 1.2 4 26 4 17 ...
 $ X4: int  64 79 70 50 36 48 92 75 23 48 ...
 $ X5: int  2 8 10 25 40 8 10 1 15 4 ...
 $ X6: int  46 67 39 58 64 55 44 44 47 36 ...
 $ Y : num  NA NA NA NA NA NA NA NA NA NA ...
> #
> #  Medidas estadísticas de base de datos con datos completos y datos faltantes
> apply(Datos, 2, mean)
      X1       X2       X3       X4       X5       X6        Y
164.43269  101.91629  12.83125  41.06731  14.29487  50.07853       NA
> apply(Datos, 2, sd)
      X1       X2       X3       X4       X5       X6        Y
135.06108 342.62989  18.57334  19.51186  12.87333  13.33857       NA
> summary(Datos$Y); sd(Datos$Y, na.rm=TRUE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      5     220     640    1431    2200    7250     458
[1] 1667.921
> As=(3*(mean(Datos$Y, na.rm=TRUE)-median(Datos$Y, na.rm=TRUE)))/sd(Datos$Y,
na.rm=TRUE); As
[1] 1.423599
> n=dim(Datos)[1]; n_F=countNA(Datos); n_C=n-n_F; n; n_F; n_C
[1] 1872
[1] 458
```

```
[1] 1414
> f=c(n_F,n_C); fr=round(prop.table(f)*100,1); cbind(f, fr); n
       f   fr
[1,]  458 24.5
[2,] 1414 75.5
[1] 1872
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> q()
> #
> #  Método de eliminación de datos
> Datos_F<-Datos[is.na(Datos$Y)!=0,]#  Datos_F=BD sólo con datos faltantes
> Datos_C<-Datos[is.na(Datos$Y)!=1,]#  Datos_C=BD sólo con datos completos
> #   También: Datos_C=na.omit(datos)   o Datos_C=na.exclude(Datos)
> apply(Datos_C, 2, mean)
       X1        X2        X3        X4        X5        X6         Y
 162.39604   101.60566   12.75438   39.20156   14.10184   50.31259 1431.48373
> apply(Datos_C, 2, sd)
       X1        X2        X3        X4        X5        X6         Y
 137.08039  349.71161   17.44087   18.55680   13.19201   13.69622 1667.92116
> summary(Datos_C$Y); sd(Datos_C$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5     220     640    1431    2200    7250
[1] 1667.921
> As=(3*(mean(Datos_C$Y)-median(Datos_C$Y)))/sd(Datos_C$Y); As
[1] 1.423599
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_C), 4)
       X1      X2     X3      X4      X5      X6      Y
X1  1.0000  0.1126  0.2047  0.1423 -0.0483  0.0163  0.4548
X2  0.1126  1.0000  0.0399  0.0242  0.0431  0.0337  0.0618
X3  0.2047  0.0399  1.0000  0.1481 -0.0479 -0.1634  0.2680
X4  0.1423  0.0242  0.1481  1.0000 -0.1578 -0.1705  0.2043
X5 -0.0483 0.0431 -0.0479 -0.1578  1.0000  0.5133 -0.0675
X6  0.0163 0.0337 -0.1634 -0.1705  0.5133  1.0000 -0.1323
Y   0.4548  0.0618  0.2680  0.2043 -0.0675 -0.1323  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos_C)[1]; Alfa=0.05
> LI= mean(Datos_C$Y)- qt(1-Alfa/2, n-1)*sd(Datos_C$Y)/sqrt(n); LI
[1] 1344.473
> LS= mean(Datos_C$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_C$Y)/sqrt(n); LS
[1] 1518.494
> R=LS-LI; R
[1] 174.0208
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos_C)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_C$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1608.634
> LS= sqrt((n-1)*var(Datos_C$Y)/qchisq(Alfa/2,n-1)); LS
```

```
[1] 1731.779
> R=LS-LI; R
[1] 123.1449
> #
> #  Método de imputación por la media
> Datos_M=Datos#  BD resultante por la imputación de la media
> Promedio=mean(Datos$Y,na.rm=TRUE)
> Datos_M[is.na(Datos_M$Y), 8] <- Promedio#  Imputar los datos faltantes por la media
> summary(Datos_M$Y); sd(Datos_M$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5     300    1431    1431    1500    7250
[1] 1449.472
> As=(3*(mean(Datos_M$Y)-median(Datos_M$Y)))/sd(Datos_M$Y); As
[1] 0
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_M), 4)
       X1      X2     X3      X4      X5      X6      Y
X1  1.0000  0.1279  0.2086  0.1377 -0.0586  0.0170  0.4012
X2  0.1279  1.0000  0.0551  0.0165  0.0256  0.0278  0.0549
X3  0.2086  0.0551  1.0000  0.1215 -0.0668 -0.1516  0.2187
X4  0.1377  0.0165  0.1215  1.0000 -0.1324 -0.1408  0.1689
X5 -0.0586 0.0256 -0.0668 -0.1324  1.0000  0.5042 -0.0601
X6  0.0170 0.0278 -0.1516 -0.1408  0.5042  1.0000 -0.1181
Y   0.4012  0.0549  0.2187  0.1689 -0.0601 -0.1181  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos_M)[1]; Alfa=0.05
> LI= mean(Datos_M$Y)- qt(1-Alfa/2, n-1)*sd(Datos_M$Y)/sqrt(n); LI
[1] 1365.781
> LS= mean(Datos_M$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_M$Y)/sqrt(n); LS
[1] 1497.187
> R=LS-LI; R
[1] 131.4062
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos_M)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_M$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1404.485
> LS= sqrt((n-1)*var(Datos_M$Y)/qchisq(Alfa/2,n-1)); LS
[1] 1497.457
> R=LS-LI; R
[1] 92.9716
> #
> #  Método de imputación por la mediana
> Datos_M=Datos#  BD resultante por la imputación de la mediana
> Mediana=median(Datos$Y,na.rm=TRUE)
> Datos_M[is.na(Datos_M$Y), 8] <- Mediana#  Imputar los datos faltantes por la mediana
> summary(Datos_M$Y); sd(Datos_M$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5     300     640    1238    1500    7250
[1] 1488.891
```

```
> As=(3*(mean(Datos_M$Y)-median(Datos_M$Y)))/sd(Datos_M$Y); As
[1] 1.204603
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_M), 4)
       X1      X2     X3      X4      X5      X6      Y
X1  1.0000  0.1279  0.2086  0.1377 -0.0586  0.0170  0.3845
X2  0.1279  1.0000  0.0551  0.0165  0.0256  0.0278  0.0530
X3  0.2086  0.0551  1.0000  0.1215 -0.0668 -0.1516  0.2113
X4  0.1377  0.0165  0.1215  1.0000 -0.1324 -0.1408  0.1260
X5 -0.0586 0.0256 -0.0668 -0.1324  1.0000  0.5042 -0.0646
X6  0.0170 0.0278 -0.1516 -0.1408  0.5042  1.0000 -0.1079
Y   0.3845  0.0530  0.2113  0.1260 -0.0646 -0.1079  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos_M)[1]; Alfa=0.05
> LI= mean(Datos_M$Y)- qt(1-Alfa/2, n-1)*sd(Datos_M$Y)/sqrt(n); LI
[1] 1170.351
> LS= mean(Datos_M$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_M$Y)/sqrt(n); LS
[1] 1305.331
> R=LS-LI; R
[1] 134.9799
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos_M)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_M$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1442.682
> LS= sqrt((n-1)*var(Datos_M$Y)/qchisq(Alfa/2,n-1)); LS
[1] 1538.182
> R=LS-LI; R
[1] 95.50005
> #
> #  Método de imputación Hot_Deck (reemplazo aleatorio)
> #
> set.seed(5000)#  Para tener los mismos resultados
> Imp_HD<- hotdeck(Datos, variable=c("Y"),
domain_var=c("X1","X2","X3","X4","X5","X6"), impNA=TRUE, imp_var=TRUE)
> Datos_Imp_HD<-Imp_HD#  BD resultante por la imputación de HotDeck (aleatorio)
> summary(Imp_HD$Y); sd(Imp_HD$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1      16     300    1082    1500    7250
[1] 1574.587
> As=(3*(mean(Imp_HD$Y)-median(Imp_HD$Y)))/sd(Imp_HD$Y); As
[1] 1.48897
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_Imp_HD), 4)
         X1      X2     X3      X4      X5      X6      Y   Y_imp
X1     1.0000  0.1279  0.2086  0.1377 -0.0586  0.0170  0.3589  0.0265
X2     0.1279  1.0000  0.0551  0.0165  0.0256  0.0278  0.0499  0.0016
X3     0.2086  0.0551  1.0000  0.1215 -0.0668 -0.1516  0.1985  0.0073
X4     0.1377  0.0165  0.1215  1.0000 -0.1324 -0.1408  0.0898  0.1681
```

```
X5    -0.0586 0.0256 -0.0668 -0.1324  1.0000  0.5042 -0.0657  0.0264
X6     0.0170 0.0278 -0.1516 -0.1408  0.5042  1.0000 -0.0967 -0.0308
Y      0.3589  0.0499  0.1985  0.0898 -0.0657 -0.0967  1.0000 -0.3906
Y_imp  0.0265 0.0016  0.0073  0.1681  0.0264 -0.0308 -0.3906  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= mean(Datos_Imp_HD$Y)- qt(1-Alfa/2, n-1)*sd(Datos_Imp_HD$Y)/sqrt(n); LI
[1] 1010.13
> LS= mean(Datos_Imp_HD$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_Imp_HD$Y)/sqrt(n); LS
[1] 1152.879
> R=LS-LI; R
[1] 142.7489
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_Imp_HD$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1525.718
> LS= sqrt((n-1)*var(Datos_Imp_HD$Y)/qchisq(Alfa/2,n-1)); LS
[1] 1626.714
> R=LS-LI; R
[1] 100.9967
> #
> #  Método de imputación por k vecino más cercano
> #
> #  Seleccionar las variables donantes
> round(cor(Datos_C), 4)
       X1      X2     X3      X4      X5      X6      Y
X1  1.0000  0.1126  0.2047  0.1423 -0.0483  0.0163  0.4548
X2  0.1126  1.0000  0.0399  0.0242  0.0431  0.0337  0.0618
X3  0.2047  0.0399  1.0000  0.1481 -0.0479 -0.1634  0.2680
X4  0.1423  0.0242  0.1481  1.0000 -0.1578 -0.1705  0.2043
X5 -0.0483 0.0431 -0.0479 -0.1578  1.0000  0.5133 -0.0675
X6  0.0163 0.0337 -0.1634 -0.1705  0.5133  1.0000 -0.1323
Y   0.4548  0.0618  0.2680  0.2043 -0.0675 -0.1323  1.0000
> Datos_S=Datos_C[, c("X1","X3","X4","X6","Y")]#  BD selecciona variables donantes
> #
> #   Simular una muestra de datos faltantes del 20% a partir de la base de datos completa
> Datos_A=Datos_S#  BD que contendrá una muestra aleatoria de 20% missing
> set.seed(5000)
> indice <- sample(2, nrow(Datos_A), replace = TRUE, prob = c(0.80, 0.20))
> Datos_A$Y[indice == 2] <-NA   #  Reemplazar por datos faltantes (NA)
> n=dim(Datos_A)[1]; n_F=countNA(Datos_A); n_C=n-n_F
> f=c(n_F,n_C); fr=round(prop.table(f)*100,1); cbind(f, fr); n
      f   fr
[1,]  267 18.9
[2,] 1147 81.1
[1] 1414
> #
> #  Análisis de sensibilidad para determinar el valor de k
> K=15
> Y_ECM<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
```

```
> X_k<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
> for(i in 1:K)
+ {
+ Imp_Knn<-kNN(Datos_A, metric=NULL, k=i, numFun=mean, imp_var=FALSE)
+ Datos_Imp_Knn<-Imp_Knn
+ ECM<-sqrt((sum((Datos_Imp_Knn$Y-Datos_C$Y)^2))/n_F); Y_ECM[i]=ECM; X_k[i]=i
+ cat("Valor del Error Cuadrado Medio para k= ",i,"   ;  ","   ECM =  ",ECM,"\n")
+ }
Valor del Error Cuadrado Medio para k=  1    ;     ECM =   1867.027
Valor del Error Cuadrado Medio para k=  2    ;     ECM =   1580.829
Valor del Error Cuadrado Medio para k=  3    ;     ECM =   1533.847
Valor del Error Cuadrado Medio para k=  4    ;     ECM =   1506.604
Valor del Error Cuadrado Medio para k=  5    ;     ECM =   1447.855
Valor del Error Cuadrado Medio para k=  6    ;     ECM =   1423.651
Valor del Error Cuadrado Medio para k=  7    ;     ECM =   1385.28
Valor del Error Cuadrado Medio para k=  8    ;     ECM =   1356.428
Valor del Error Cuadrado Medio para k=  9    ;     ECM =   1334.202
Valor del Error Cuadrado Medio para k=  10    ;     ECM =   1346.53
Valor del Error Cuadrado Medio para k=  11    ;     ECM =   1342.638
Valor del Error Cuadrado Medio para k=  12    ;     ECM =   1343.369
Valor del Error Cuadrado Medio para k=  13    ;     ECM =   1341.951
Valor del Error Cuadrado Medio para k=  14    ;     ECM =   1344.298
Valor del Error Cuadrado Medio para k=  15    ;     ECM =   1345.916
> plot(X_k, Y_ECM, type="b", xlab="Valores de k", ylab="Valores de ECM")
> #
> #  Método de imputación por k vecino más cercano (con la media)
> Datos_S=Datos[, c("X1","X3","X4","X6","Y")]#  BD selecciona variables donantes
> Imp_Knn<-kNN(Datos_S, metric=NULL, k=9, numFun=mean, imp_var=FALSE)
> Datos_Imp_Knn<-Imp_Knn#  BD resultante por la imputación de k vecino más cercano
> summary(Imp_Knn$Y); sd(Imp_Knn$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0   280.0   857.2  1425.0  2076.9  7250.0
[1] 1510.955
> As=(3*(mean(Imp_Knn$Y)-median(Imp_Knn$Y)))/sd(Imp_Knn$Y); As
[1] 1.12741
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_Imp_Knn), 4)
       X1      X3      X4      X6       Y
X1 1.0000  0.2086  0.1377  0.0170  0.4799
X3 0.2086  1.0000  0.1215 -0.1516  0.2480
X4 0.1377  0.1215  1.0000 -0.1408  0.2054
X6 0.0170 -0.1516 -0.1408  1.0000 -0.1387
Y  0.4799  0.2480  0.2054 -0.1387  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= mean(Datos_Imp_Knn$Y)- qt(1-Alfa/2, n-1)*sd(Datos_Imp_Knn$Y)/sqrt(n); LI
[1] 1356.498
> LS= mean(Datos_Imp_Knn$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_Imp_Knn$Y)/sqrt(n); LS
[1] 1493.479
> R=LS-LI; R
```

```
[1] 136.9802
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_Imp_Knn$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1464.06
> LS= sqrt((n-1)*var(Datos_Imp_Knn$Y)/qchisq(Alfa/2,n-1)); LS
[1] 1560.976
> R=LS-LI; R
[1] 96.91524
> #
> #  Método de imputación por k vecino más cercano (con la mediana)
> Imp_Knn<-kNN(Datos_S, metric=NULL, k=9, numFun=median, imp_var=FALSE)
> Datos_Imp_Knn<-Imp_Knn#  BD resultante por la imputación de k vecino más cercano
> summary(Imp_Knn$Y); sd(Imp_Knn$Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0   250.0   644.5  1335.2  1943.0  7250.0
[1] 1537.851
> As=(3*(mean(Imp_Knn$Y)-median(Imp_Knn$Y)))/sd(Imp_Knn$Y); As
[1] 1.347326
> hist(Datos$Y,  breaks=10, freq=NULL, main=" ", xlab="Ingreso de los hogares",
ylab="Número de hogares")
> round(cor(Datos_Imp_Knn), 4)
       X1      X3      X4      X6       Y
X1 1.0000  0.2086  0.1377  0.0170  0.4833
X3 0.2086  1.0000  0.1215 -0.1516  0.2533
X4 0.1377  0.1215  1.0000 -0.1408  0.1651
X6 0.0170 -0.1516 -0.1408  1.0000 -0.1222
Y  0.4833  0.2533  0.1651 -0.1222  1.0000
> #
> #  Cálculo del intervalo de confianza del 95% para la media del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= mean(Datos_Imp_Knn$Y)- qt(1-Alfa/2, n-1)*sd(Datos_Imp_Knn$Y)/sqrt(n); LI
[1] 1265.453
> LS= mean(Datos_Imp_Knn$Y)+ qt(1-Alfa/2, n-1)*sd(Datos_Imp_Knn$Y)/sqrt(n); LS
[1] 1404.872
> R=LS-LI; R
[1] 139.4185
> #
> #  Cálculo del intervalo de confianza del 95% para la desviación estándar del ingreso
> n=dim(Datos)[1]; Alfa=0.05
> LI= sqrt((n-1)*var(Datos_Imp_Knn$Y)/qchisq(1-Alfa/2,n-1)); LI
[1] 1490.122
> LS= sqrt((n-1)*var(Datos_Imp_Knn$Y)/qchisq(Alfa/2,n-1)); LS
[1] 1588.762
> R=LS-LI; R
[1] 98.64041
> #
> #  Comparación de los métodos de imputación con el ECM
> #
> #   Simular una muestra de datos faltantes del 10% a partir de la base de datos completos
> Datos_A=Datos_C#  BD que contendrá una muestra aleatoria de 10% missing
```

```
> set.seed(5000)
> indice <- sample(2, nrow(Datos_A), replace = TRUE, prob = c(0.90, 0.10))
> Datos_A$Y[indice == 2] <-NA
> n=dim(Datos_A)[1]; n_F=countNA(Datos_A); n_C=n-n_F
> f=c(n_F,n_C); fr=round(prop.table(f)*100,1); cbind(f, fr); n
      f   fr
[1,]  126  8.9
[2,] 1288 91.1
[1] 1414
> #
> #  Calcular el ECM entre valores imputados y observados para cada uno de los métodos
> #  Método de imputación por la media
> Datos_M=Datos_A
> Promedio=mean(Datos_M$Y,na.rm=TRUE)
> Datos_M[is.na(Datos_M$Y), 8] <- Promedio
> ECM<-sum((Datos_M$Y-Datos_C$Y)^2);sqrt(ECM/n_F)
[1] 1504.549
> cor(Datos_M$Y, Datos_C$Y)
[1] 0.9630815
> #  Método de imputación por la mediana
> Datos_M=Datos_A
> Mediana=median(Datos_M$Y,na.rm=TRUE)
> Datos_M[is.na(Datos_M$Y), 8] <- Mediana
> ECM<-sum((Datos_M$Y-Datos_C$Y)^2);sqrt(ECM/n_F)
[1] 1619.906
> cor(Datos_M$Y, Datos_C$Y)
[1] 0.9577189
> #  Método de imputación HotDeck (aleatorio)
> Imp_HD<- hotdeck(Datos_A, variable=c("Y"),
domain_var=c("X1","X2","X3","X4","X5","X6"), impNA=TRUE, imp_var=TRUE)
> Datos_Imp_HD<-Imp_HD
> ECM<-sum((Datos_Imp_HD$Y-Datos_C$Y)^2);sqrt(ECM/n_F)
[1] 1963.671
> cor(Datos_Imp_HD$Y, Datos_C$Y)
[1] 0.9401904
> #  Método de imputación por k vecino más cercano (con la media)
> Datos_S=Datos_A[, c("X1","X3","X4","X6","Y")]
> Imp_Knn<-kNN(Datos_S, metric=NULL, k=9, numFun=mean, imp_var=FALSE)
> Datos_Imp_Knn<-Imp_Knn
> ECM<-sum((Datos_Imp_Knn$Y-Datos_C$Y)^2);sqrt(ECM/n_F)
[1] 1412.6
> cor(Datos_Imp_Knn$Y, Datos_C$Y)
[1] 0.9676415
> #  Método de imputación por k vecino más cercano (con la mediana)
> Imp_Knn<-kNN(Datos_S, metric=NULL, k=9, numFun=median, imp_var=FALSE)
> Datos_Imp_Knn<-Imp_Knn
> ECM<-sum((Datos_Imp_Knn$Y-Datos_C$Y)^2);sqrt(ECM/n)
[1] 444.3896
> cor(Datos_Imp_Knn$Y, Datos_C$Y)
[1] 0.9641368
>
```