UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

ESTIMACIÓN DE DATOS FALTANTES A TRAVÉS DE REDES NEURONALES, UNA COMPARACIÓN CON MÉTODOS SIMPLES Y MÚLTIPLES

KELLY JULIANA BÁEZ VERGARA

DIRIGIDO POR:

ROBERT ROMERO

OPCIÓN DE GRADO - TESIS

BOGOTÁ D.C.

2022

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ESQUEMAS

1. INTRODUCCIÓN

Encontrarse con valores faltantes es un problema muy frecuente en la data, la ausencia de valores en los campos puede llegar a causar problemas de sesgo en las estimaciones, así como alterar la distribución y varianza de los datos. El proceso de imputación está incluido dentro de la etapa de preparación de datos, que es el paso que se realiza luego de llevar a cabo el entendimiento de los mismos. Aquí, los valores faltantes son reemplazados por valores estimados conocidos, esto con el fin de obtener un conjunto de datos completo, al cual se le pueda aplicar diversas técnicas estadísticas. La importancia de encontrar un correcto método de imputación es esencial, ya que trabajar con datos equivocados, implicaría tener alteraciones en el resultado final de las estimaciones.

De acuerdo con algunos autores, las técnicas simples de imputación pueden tener algunas ventajas sobre las múltiples, ya que se dice que hay menos riesgo de pérdida de eficiencia en comparación con las técnicas múltiples (Little y Rubin, 2002). Los métodos aleatorios pueden dar mayor variabilidad respecto a las determinísticas, ya que estas últimas pueden subestimar más la varianza. Sin embargo, su ventaja está en que estás tienden a generar estadísticas más precisas que las técnicas aleatorias. La imputación múltiple trabaja creando varias estimaciones sobre la data, que luego deben ser combinadas para obtener el valor estimado, ayudando a reducir el sesgo, aumentar la precisión y así se obtener estadísticas muy sólidas.

El presente trabajo, propone realizar la estimación de datos faltantes de diversas técnicas tanto simples como múltiples, y así comparar los resultados obtenidos en los diferentes escenarios. Dentro de los métodos utilizados se propone el uso de redes neuronales, ya que su metodología brinda un enfoque alternativo con mayor eficiencia que los convencionales.

Las redes neuronales están asociadas a la estructura de un cerebro humano, ya que trabajan de manera análoga a las funciones más elementales de la neurona biológica. Una red neuronal se compone de unidades que se encuentran en una serie de capas, cada una de las cuales se conecta a las capas de cada lado. Algunas se reconocen como capas de entrada,

que están diseñadas para recibir diversas formas de información del mundo exterior que la red intentará aprender, reconocer o procesar de otra manera. En el lado opuesto, se encuentran las capas de salida, que dan el resultado final de todo el procesamiento de la red. Entre las capas de entrada y de salida se encuentran las capas ocultas. La mayoría de las redes neuronales están totalmente conectadas , lo que significa que cada unidad oculta y cada unidad de salida está conectada a cada unidad en las capas a ambos lados. Las conexiones se representan a través de pesos, que pueden tomar valores positivos o negativos. Cuanto mayor sea el peso, más influencia tiene una unidad sobre otra. (Woodford, C. 2021).

2. OBJETIVO GENERAL

Comparar la eficiencia de la imputación de valores perdidos realizada por diferentes métodos frente al método de redes neuronales, con miras de identificar su eficiencia respecto a los demás.

3. OBJETIVOS ESPECÍFICOS

3.1. Identificar la eficiencia del modelo de redes neuronales a diferentes tasas de valores perdidos.

3.2. Comparar el impacto de la imputación de datos faltantes con procesos de transformación de variables.

3.3. Determinar si el método de imputación implementado genera algún cambio significativo en términos de la distribución de los datos estimados en comparación con la data original.

4. PLANTEAMIENTO DEL PROBLEMA

Para el presente trabajo, se propone comparar la eficiencia de diferentes métodos de imputación sobre diferentes tipos de datos que cuentan con porcentajes de datos faltantes determinados. Para ellos se plantean las siguientes consideraciones.

4.1. La información con la cual se desarrolla este trabajo proviene de la encuesta multipropósito del año 2017, publicada por el DANE.

4.2. Dentro de las variables consideradas en los métodos de imputación se destacan:

a. Valor mensual pagado por concepto de arriendo.
b. Valor mensual gastado en mercado.
c. Valor mensual gastado en servicios públicos.
d. Valor mensual gastado en productos de aseo personal
e. Valor mensual gastado en productos de aseo personal para el hogar.

4.3. Se establecen para cada una de las variables anteriormente mencionadas porcentajes de valores faltantes, los cuales se agrupan en 2%, 5%, 10% y 20%.

5. ANTECEDENTES

La ausencia de valores en conjuntos de datos ha sido un problema recurrente al cual se tiene que enfrentar los analistas de datos, especialmente cuando se cuenta con un conjunto de datos pequeño y una gran proporción de faltantes, pero con el paso de los años, se han venido estudiando diversas técnicas para "llenar" estos vacíos, ya que estos no deben ser ignorados, sino que requieren de un correcto tratamiento. Sin embargo, esta labor se complica cuando son múltiples variables las que tienen esta problemática.

Gracias a los avances computacionales de las últimas décadas, se han planteado diferentes formas de estudiar los datos faltantes multivariantes, no obstante su aplicación debe ser adecuada según el contexto y necesidad, con el fin de evitar que se presenten sesgos en las estimaciones, relación entre las variables, entre otros.

Son varias las razones por las cuales un campo cuenta con ausencia de información de una variable, y es importante identificar si se trata de una pérdida total o parcial, para así realizar el tratamiento adecuado. En la práctica, comúnmente se habla de que una pérdida de datos entre el 1% y 20% es manejable, sin embargo gracias a los avances en métodos múltiples, se habla de que estos métodos aún llegan a tener gran eficiencia hasta en tasas del 50%. De igual manera, se habla que tasas de faltantes menores al 10%, tienen unos resultados similares independiente del método utilizado.

5.1. Tipos de no respuesta

En las encuestas de los estudios estadísticos, es común encontrar algunos campos de no respuesta en las variables, la cual se puede presentar de dos maneras:

a. **No respuesta total**: es cuando no se encuentra ningún dato en la unidad de muestra. Por ejemplo, al momento de aplicar la encuesta a un hogar previamente seleccionado, no es posible encuestar ya que no había alguien encargado.
b. **No respuesta parcial:** se describe como la ausencia de respuesta en algunas variables, pero no ausencia completa en todo su registro. (Useche y Mesa, 2006).

5.2. Tipos de datos faltantes

Dentro de los tipos de no respuesta, es importante identificar que tipo de patrón describen los datos faltantes, ya que esto puede influir en la selección del método de imputación

a. **MCAR** (Missing Completely At Random) este tipo de datos no muestran una relación entre la ausencia y las covariables observadas o no observadas.

b. **MAR** (Missing At Random) si bien los faltantes también son aleatorios, estos dependen de las variables observadas. Por ejemplo, un grupo de personas con un nivel socioeconómico alto, estarían menos dispuestos a entregar información salarial. Es decir, los faltantes no se deben a los valores que no se observan. MCAR implica MAR pero no el caso contrario.

c. **MNAR** (Missing Not At Random) acá se encuentran los datos que faltan que dependen de otros faltantes, por ejemplo, un dispositivo que mide alguna respuesta superior al 0.5, los valores que estén por debajo de este valor, serán vacíos.

5.2.1. Test de little

En la prueba de Little de MCAR los datos y_i (i = 1, 2, ..., n) se modelan como normal multivariante p-dimensional con vector medio μ y matriz de covarianza Σ, con parte de los componentes en y_i falta. Cuando no se satisface la normalidad, la prueba de Little todavía funciona en el sentido asintótico para los vectores aleatorios cuantitativos y_i, pero no es adecuada para las variables categóricas. (Li, 2013).

La estadística de prueba de χ2 de Little para MCAR tiene la siguiente forma:

$$d_0^2 = \sum_{j=1}^{J} n_j \, (\bar{y}_{oj} - \mu_{oj}) \; {}^{T}\Sigma_{oj}^{-1} \; (\bar{y}_{oj} - \mu_{oj}) \qquad (5.2.1.1.)^{1}$$

La idea es que si los datos son MCAR, entonces condicional al indicador faltante r_i, se cumple la siguiente hipótesis nula:

$$H_0 = y_{o,i} | r_i \sim N(\mu_{oj}, \Sigma_{oj}) \qquad si \;\; i \in I_j, 1 \leq j \leq J \qquad (5.2.1.2.)^{2}$$

Donde μ_{oj} Es un subvector del vector de medias μ.

En cambio, si la hipótesis no es cierta, entonces la condicional al indicador faltante r_i, se espera que las medias de las y observadas varíen a través de diferentes patrones, lo que implica (Li, 2013).

$$H_1 = y_{o,i} | r_i \sim N(\upsilon_{oj}, \Sigma_{oj}) \qquad si \;\; i \in I_j, 1 \leq j \leq J. \qquad (5.2.1.3.)^{3}$$

[1] Li, C. (2013). Little's test of missing completely at random. The Stata Journal, 13(4), 795–809.
[2] Li, C. (2013). Little's test of missing completely at random. The Stata Journal, 13(4), 795–809.
[3] Li, C. (2013). Little's test of missing completely at random. The Stata Journal, 13(4), 795–809.

5.2.1.1. Test de little resultados

A través del algoritmo EM (Expectation-Maximization) de SPSS, se lleva a cabo el análisis de las variables que poseen valores perdidos, encontrando los siguientes valores para la prueba MCAR de little:

$H_0 =$ Los datos son MCAR

$H_1 =$ Los datos no son MCAR

Datos	**Chi-cuadrado**	**DF**	**Sig. (p-valor)**
Conjunto 2%	82915687467002	2815	0
Conjunto 5%	3206,498	3813	1
Conjunto 10%	220698843219819	4457	0
Conjunto 20%	0	4667	1

Tabla 1: Resultado test de little. Elaboración propia.

Con un nivel de significancia del 5%, no hay evidencia estadística suficiente para rechazar H_0 para los conjuntos de datos que poseen valores faltantes al 5% y al 20%. Mientras que para los conjuntos de datos con faltantes al 2% y al 10% se rechaza H_0.

5.3. Técnicas de imputación seleccionadas

Dentro del trabajo desarrollado, se proponen métodos de imputación, los cuales se describen:

5.3.1. Imputación por mediana

Este método es uno de los más utilizados en la práctica, debido a su fácil implementación. Este aplica sólo a variables tipo numérico ya sea continuas o discretas, donde básicamente calcula el valor de la mediana en el conjunto de datos disponible, y este valor luego será reemplazado en los valores faltantes.

Valor arriendo
450.000
1.254.000
678.900
1.080.900
1.294.040
789.000
1.507.180
1.587.000
965.000
1.150.000

Mediana
1.115.450

Valor arriendo imputado
450.000
1.115.450
1.254.000
678.900
1.080.900
1.115.450
1.294.040
789.000
1.507.180
977.950
1.115.450
965.000
1.115.450

Esquema 1: Imputación mediana. Elaboración propia

Esta técnica tiene las siguientes consideraciones:

- Si la variable sigue una distribución normal, la media y la mediana son aproximadamente iguales.
- Si la variable tiene una distribución asimétrica, entonces la mediana es una mejor representación.
- Dentro de sus limitaciones, se destaca que puede distorsionar la distribución y la varianza de la variable original, y esto puede aumentar cuanto mayor sea el porcentaje de valores perdidos.

5.3.2. Imputación multivariante mediante ecuaciones encadenadas

Con su siglas en inglés MICE, asume que los datos faltantes son Missing at Random (MAR), es decir, que la probabilidad de que falte un valor depende solo del valor observado y se puede predecir usándolos. Imputa datos variable por variable especificando un modelo de imputación por variable. (Azur, Stuart, Frangakis and Leaf, 2011).

Algunos modelos de imputación de datos, asume que todas las variables que componen la data presentan una distribución conjunta, la ventaja es usar MICE es que su algoritmo ejecuta modelos de regresión mediante los cuales cada variable con faltantes se modela de forma condicionada a las otras variables de los datos, otras palabras, cada variable se moldea de acuerdo con su distribución.

Suponga que la data es $X = (X^{miss}, X^{obs})$ donde X^{miss} representa las variables al menos tiene un valor faltante y X^{obs} incluye las columnas de X completamente observadas. X_j es la j-ésima columna de X_j^{miss} es el valor faltante un la j-ésima columna de X^{miss}. Z es la matriz imputada de X y k es el número parcial de variables observada. Se define X_{-j}^{miss} igual a la matriz X con su j-ésima columna eliminada. Así, tendremos el siguiente algoritmo para mostrar la implementación de MICE estándar (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021):

1. Para rellenar los valores inicial de los valores perdidos, se define una matriz Z igual a X^{obs}; para cada X_j^{miss} valores son rellenados con extracciones aleatorias de la distribución predictiva condicionada en Z, y se asocia la versión imputada de X_j^{miss} a Z antes de incrementar *j*. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).
2. Para $j = 1, \ldots k$, reemplaza los valores faltantes de X_j con extracciones aleatorias de la distribución predictiva condicional en X_{-j}^{miss}. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).

3. Se repiten los pasos 1 y 2 para un número de iteraciones. Este procedimiento se realiza para cada variable con al menos un valor perdido, lo que da lugar a un conjunto de datos completo. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).
4. Repita los pasos 1 a 3 un número de veces (*M*), dando como resultado *M* conjuntos de datos imputados que se analiza por separado. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).

5.3.3. Imputación multivariante mediante ecuaciones encadenadas con bosques aleatorios (miceforest)

La imputación de random forest es una técnica no paramétrica que presenta una alternativa desarrollada recientemente a los procedimientos MICE estándar. Random forest es una extensión de los árboles de clasificación y regresión y utilizan un enfoque de división binaria que subdivide los datos en función de los valores de las variables predictoras, construyendo muchos árboles cada vez que varían las muestras y los predictores.

La creación de múltiples conjuntos de datos con diferentes valores imputados le permite hacer dos tipos de inferencia:

- **Distribución del valor imputado:** se puede construir un perfil para cada valor imputado, lo que le permite hacer declaraciones sobre la probable distribución de ese valor.
- **Distribución de predicciones de modelos:** con varios conjuntos de datos, puede crear varios modelos y crear una distribución de predicciones para cada muestra. Aquellas muestras con valores imputados que no pudieron imputarse con mucha confianza tendrían una mayor varianza en sus predicciones.

Su proceso se da a través de los siguientes pasos:

5.3.3.1. Random Forest

Para b=1 a B:

1. Extrae una muestra Z^* de tamaño N de la data de entrenamiento

2. Crece un árbol random forest T_b para la muestra extraída, repitiendo recursivamente los siguientes pasos para cada nodo terminal de el árbol, hasta que se alcance el tamaño mínimo de nodo n_{min} :

a. Seleccionar m variables aleatoriamente de entre las p variables.

b. Elegir la mejor variable/punto de corte entre las m.

c. Dividir el nodo en dos nodos hijos.

Salida del conjunto de árboles $\{T_b\}_1^B$. Para hacer una predicción en un nuevo conjunto x: $\widehat{f}_{rf}^{B}(x) = \frac{1}{B}\Sigma_{b=1}^{B} T_b(x)$, (Hastie, Tibshirani, & Friedman, 2009).

5.3.3.2. MICEFOREST

1. Para los sujetos en X_j^{miss}, encuentra el nodo terminal; este termina de acuerdo al árbol ajustado en el paso anterior, y un valor observado en X_j^{miss} se selecciona de forma aleatoria del subconjunto de este este nodo y es usado para la imputación. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).
2. Repita los pasos 2 y 3 durante un número de iteraciones. Este procedimiento se realiza para cada variable con al menos un valor perdido, lo que da lugar a un conjunto de datos completo. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).
3. Repita los pasos 1-4 un número de veces (M), dando como resultado M conjuntos de datos imputados. (Javadi, Bahrampour, Saber, Garrusi, & Baneshi, 2021).

5.3.2. Redes neuronales: MIDAS

Multiple Imputation with Denoising Autoencoders por sus siglas MIDAS, emplea redes neuronales no supervisadas conocidas como autoencoders de eliminación de ruido o denoising autoencoders (DA), que trabajan en la reducción de dimensionalidad y reconstruyen el conjunto de datos. Estas redes constan de una serie de funciones no lineales anidadas que normalmente se representan como nodos interconectados organizados en capas.(Lall and Robinson, 2021). Los datos de entrada se introducen en la red a través de una capa de entrada, los nodos los procesan en una o más capas ocultas y los devuelven a través de los nodos en una capa de salida, que se puede representar :

$$y^{(h)} = \sigma(W^{(h)}y^{(h-1)} + b^{(h)}) \quad (5.3.2.1)^4$$

donde $y^{(h)}$ es un vector de salidas de la capa h, $W^{(h)}$ es una matriz de pesos que conecta los nodos en la capa $h - 1$ con los nodos en la capa h, b Es un vector de sesgos para la capa h y es una función de activación no lineal. La introducción de la no linealidad en el modelo permite que las redes neuronales aprendan de manera eficiente formas funcionales complejas con pocas capas ocultas (Lall and Robinson, 2021). Este modelo puede ser generalizado a un número arbitrario de capas ocultas H:

$$y^{(h)} = \phi\,(W^{(h)}\,[\ldots\,[\sigma(W^{(2)}[\sigma(W^{(1)}x + b^{(1)})] + b^{(2)})]\ldots] + b^{(H)}) \quad (5.3.2.2)^5$$

donde x es un vector de entradas y ϕ Es una función de activación de capa final que devuelve salidas con la distribución adecuada.

5.3.2.1 Autoencoders clásicos

Los denoising autoencoders tiene nueva adaptabilidad en la imputación de observaciones faltantes. Estos con una extensión de los autoencoders clásicos siendo una herramienta bien establecida para la reducción de la dimensionalidad en el aprendizaje automático, estos constan de dos partes. Primero, un encoder asigna de manera determinista un vector de entrada x a una representación de menor dimensión y lo comprime a través de una serie de capas ocultas que se reducen (Lall and Robinson, 2021):

$$f_{\theta}(x) = \sigma(W^{(B)}[\ldots\,[\sigma(W^{(2)}[\sigma(W^{(1)}x + b^{(1)})] + b^{(2)})]\ldots] + b^{(B)}) \quad (5.3.2.3)^6$$

En segundo lugar, un decoder regresa un vector z reconstruido con la misma distribución de probabilidad y dimensiones que x pasándolo a través de una serie paralela de capas ocultas en expansión que culminan en la capa de salida, (Lall and Robinson, 2021):

[4] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.
[5] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.
[6] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.

$$z = g_{\theta'}(y) = \phi\,(W^{(H)'}[\ldots[\sigma(W^{(B+2)'}[\sigma(W^{(B+1)'}y + b^{(B+1)'})] + b^{(B+2)'})]\ldots] + b^{(H)'})$$
(5.3.2.4)[7]

Para mapear z lo más cerca posible de x, los pesos se ajustan mediante retropropagación para minimizar una función de pérdida $L(x, y)$. Este proceso produce una representación latente que captura los ejes clave de variación en x de manera similar al análisis de componentes principales. (Lall and Robinson, 2021).

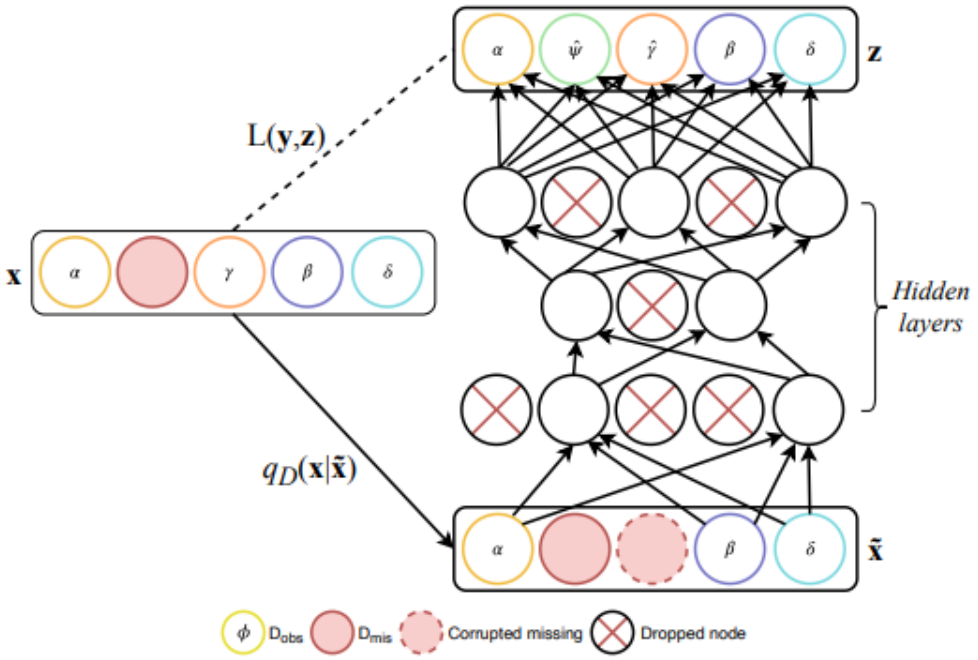
5.3.2.2. Denoising autoencoders

Los DA se desarrollaron para evitar que los autoencoders aprendan una representación idéntica de la entrada (la función de identidad) al tiempo que les permite extraer características más sólidas de los datos. Logran estos beneficios al corromper parcialmente las entradas a través de la inyección de ruido estocástico: $x \to \bar{x} \sim qD(x|\bar{x})$. Luego, la entrada corrupta se asigna a una representación oculta $y = f_{\theta}(\bar{x})$, a partir de la cual se reconstruye una versión limpia o "sin ruido" $z = g_{\theta'}(y)$. (Lall and Robinson, 2021).

5.3.2.3. MIDAS: implementación

MIDAS modifica el modelo DA estándar de dos maneras clave. Primero, como parte del proceso de corrupción inicial, obliga a todos los valores faltantes, además de un subconjunto aleatorio de entradas, a inicial en 0. La tarea del DA es predecir los valores corruptos que faltaban originalmente ($\bar{x}_{miss}$) y observada originalmente ($\bar{x}_{obs}$) usando una función de pérdida que solo incluye a esta última. En segundo lugar, para reducir aún más el riesgo de sobreajuste, MIDAS regulariza la DA con la técnica complementaria de abandono, que implica la eliminación aleatoria de nodos en las capas ocultas de una red durante el entrenamiento. (Lall and Robinson, 2021).

[7] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.



Esquema 2: Red Neuroral MIDAS. (Lall and Robinson, 2021).

El entrenamiento de abandono procede mediante el muestreo de un número arbitrario de redes "delgadas", con un conjunto diferente de nodos descartados en cada iteración. Para producir imputaciones múltiples, MIDAS r muestrea M redes reducidas. La implicación es que MIDAS postula no una distribución conjunta de los datos sino una distribución sobre posibles funciones que describen los datos. MIDAS puede capturar una clase más amplia de distribuciones conjuntas que los enfoques existentes para imputación múltiple sin hacer suposiciones paramétricas adicionales.(Lall and Robinson, 2021).

El encoder de un DA generador de imputación entrenado con abandono, una red MIDAS, puede describirse como:

$$\bar{y} = f_{\theta}(\bar{x}) = \sigma(W^{(B)}v^{(B)}[\ldots[\sigma(W^{(2)}v^{(2)}[\sigma(W^{(1)}\bar{x} + b^{(1)})] + b^{(2)})]\ldots] + b^{(B)})$$

(5.3.2.3.1)[8]

El decodificador se convierte en:

$$z = g_{\theta'}(\bar{y}) = \phi(W^{(H)'}[\ldots[\sigma(W^{(B+2)'}[\sigma(W^{(B+1)'}\bar{y} + b^{(B+1)'})] + b^{(B+2)'})]\ldots] + b^{(H)'})$$

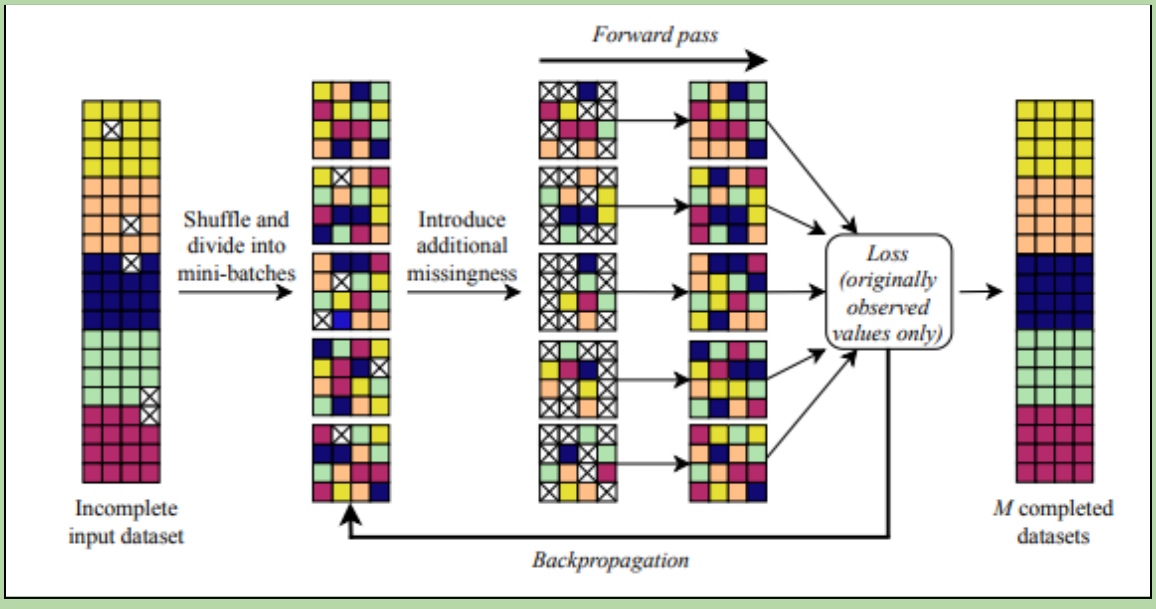
[8] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.

(5.3.2.3.1)[9]

donde g y z representa un vector completamente observado que contiene predicciones de ($\bar{x}_{obs}$)y ($\bar{x}_{miss}$). Para producir un conjunto de datos completo, las predicciones de ($\bar{x}_{miss}$) se sustituyen por x_{miss} en D.(Lall and Robinson, 2021).

5.3.2.4. Algoritmo

El algoritmo toma un conjunto de datos incompleto D como entrada y devuelve M conjuntos de datos completos. El algoritmo procede en tres etapas. En la primera etapa, los datos de entrada D se preparan para el entrenamiento. Las variables categóricas están codificadas "one-hot" y las variables continuas se reescalan entre 0 y 1 para mejorar la convergencia. Además, se construye una matriz indicadora de faltantes R para D, todos los elementos ausentes se establecen en 0. A continuación, se inicializa un DA de acuerdo con las dimensiones de DLa arquitectura predeterminada es una red de tres capas con 256 nodos por capa.



Esquema 3: Algoritmo red neuronal MIDAS. (Lall and Robinson, 2021).

Finalmente, una vez que se completa el entrenamiento, la totalidad de D se pasa al DA, que intenta reconstruir todos los valores dañados (es decir, originalmente observados y originalmente faltantes). Luego, se construye un conjunto de datos completo al reemplazar

[9] Lall, R., & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.

los faltantes con predicciones de los valores originalmente faltantes de la salida de la red. Esta etapa se repite M veces.(Lall and Robinson, 2021).

6. ENTENDIMIENTO DE DATOS

6.1. Fuentes de información

Las variables utilizadas son obtenidas de la Encuesta Multipropósito del año 2017, principalmente de los siguientes capítulos:

- Capítulo A. Identificación
- Capítulo B: Datos de la vivienda y su entorno
- Capítulo C: Condiciones habitacionales del hogar
- Capítulo D: Servicios públicos domiciliarios y de TI
- Capítulo M1: Gastos en alimentos y bebidas no alcohólicas de los hogares

6.2. Selección de variables

Para realizar el presente trabajo se toman variables de tipo cuantitativo que describen algunos gastos que comúnmente tienen los hogares. Estas están presentadas en términos de pesos colombianos y corresponde a aquellas que se van a estimar con los métodos propuestos:

Variables	**Descripción**
NHCCP10	Valor mensual pagado de arriendo o leasing
MERCADO	Monto total mensual gastado en alimentos
SERVICIOS_PUB	Monto total mensual por concepto de agua, luz, gas e internet
NHCMP7AA	Valor mensual destinado a comprar de producto de aseo para el hogar
NHCMP7BA	Valor mensual destinado a comprar de producto de aseo personal

Tabla 2: Variables a imputar. Elaboración propia

No obstante, se cuenta con variables auxiliares, las cuales también permiten hacer un acercamiento sobre el tipo de hogar y que a su vez son indispensables para las estimaciones. **Ver anexo 1**.

6.3. Descripción de la información

Del total de variables que se muestran en el diccionario, se crearon grupos de variables que comparten características similares y describen un mismo objeto, las cuales se muestran así: En el módulo arriendo, se muestran los hogares que pagan arriendo o cuota de leasing habitacional, en características del hogar se muestran algunas variables que permiten idealizar cómo está conformado el espacio donde habita el hogar y si este cuenta con algunos objetos básicos de una vivienda promedio.

Para servicios públicos y de TI se seleccionaron los básicos del hogar, contando el internet en este grupo, dado que este ha ganado gran relevancia para el desarrollo de distintas actividades diarias y se considera un importante indicador del estatus del hogar para así poderlo caracterizar. Las características de vivienda por su parte, se relacionan más con el entorno y características generales con las que cuenta la vivienda.

Por su parte se tiene un listado de problemas existentes de infraestructura y entorno los cuales son dicotomizados de tal manera que se identifique el hogar como si no tuviera el problema descrito. Finalmente, se cuenta con un módulo de gastos, que intenta hacer un reconocimiento sobre la distribución de estos en cada hogar, tomando gastos básicos como alimentación, así como transporte, y en productos de aseo personal y de sostenimiento del hogar.

7. PREPARACIÓN DE DATOS

7.1. Selección de datos

Teniendo en cuenta el objetivo del presente trabajo, se decide trabajar con directorios de Bogotá que estén conformados con un solo hogar, ya que trabajar con aquellos que cuentan con más de un hogar, no permitiría utilizar el valor real de algunas variables tales como servicios públicos, valor pagado de arriendo, entre otras, dado que sería necesario promediar o transformar para obtener una única respuesta por vivienda, y se pueden sesgar los resultados esperados.

7.2. Transformación y tratamiento de datos

Dentro de los procedimientos realizados a las variables utilizadas se destacan dos procesos aplicados a la data, donde se realiza una transformación binaria a las variables discretas, y de también se realiza una limpieza de campos que no corresponden a la descripción de la variable.

7.2.1. Binarización

Sea x una variable categórica con respuesta de '1' y '2', siendo '1' la afirmación al campo y '2' la negación:

$$x = \begin{cases} x = 1 & X = '2' \\ x = 0 & e.o.c. \end{cases}$$

7.2.2. Transformación

Se hace una limpieza de los valores que se ubican en la cola izquierda de la distribución de las variables a estimar, con el fin de obtener una muestra más homogénea y con valores más relacionados a un contexto real:

- **S**i $NHCCP10$ < Percentil 1 entonces omite.
- **S**i $MERCADO$ < Percentil 5 entonces omite.

- **S**i $SERVICIOS\_PUB$ < Percentil 1 entonces omite.
- **S**i $NHCMP7AA$ < Percentil 10 entonces omite.
- **S**i $NHCMP7BA$ < Percentil 10 entonces omite.

8. DESCRIPTIVO DE LOS DATOS

Para el presente trabajo se cuenta con un total de 17.849 registros, los cuales tomando información de las 21 localidades de la ciudad de Bogotá, siendo la población más representativa la localidad de Kennedy y Suba con un 11.61% y 10.38% respectivamente. Localidad como Sumapaz y otras rurales no representan ni un 1% respecto al total.

Sin embargo, pese a no ser una localidad con gran número de población, Chapinero se ubica en el primer lugar en términos de valor de arriendo, con un valor de $1.495.814 seguida, no por mucha diferencia, por la localidad de Teusaquillo. Por su parte, Usme y Ciudad Bolívar son aquellas que cuentan con menor valor promedio pagado por concepto de arriendo, con $365.552 y $440.554 respectivamente.

En términos de valor mensual promedio destinado a cubrir los gastos relacionados a mercado del hogar, se evidencia que la localidad de Sumapaz tiene el valor más alto con $487.949, mientras que Usme, siendo la última en la lista, presenta un valor de $235.064. Para la variable que describe el valor mensual promedio gastado en el pago de servicios públicos, se muestra que Usaquen y Chapinero tienen un valor de $321.923 y $279.654, siendo las localidades que encabezan la lista con los precios más altos, respectivamente.

Localidad	Num. Hogares	% Hogares	Prom. NHCCP10	Prom. MERCADO	Prom. SERVICIOS_PUB	Prom. NHCMP7AA	Prom. NHCMP7BA
ANTONIO NARIÑO	504	2,82%	792.635	341.160	226.984	75.770	71.211
BARRIOS UNIDOS	470	2,63%	1.207.796	369.912	275.679	84.412	76.594
BOSA	940	5,27%	464.992	265.705	159.448	56.746	53.387
CANDELARIA	225	1,26%	976.615	305.918	157.332	70.458	65.489
CHAPINERO	471	2,64%	1.495.814	315.894	279.654	86.006	73.822
CIUDAD BOLIVAR	1096	6,14%	440.554	243.443	121.543	46.503	45.035
ENGATIVA	1395	7,82%	776.166	296.599	195.398	75.148	70.036
FONTIBON	1576	8,83%	979.326	349.777	232.835	81.539	71.673
KENNEDY	2073	11,61%	610.784	271.016	177.025	61.558	58.909
LOS MÁRTIRES	508	2,85%	757.907	268.555	252.839	68.779	64.315
LOCALIDAD RURAL	43	0,24%	670.698	267.326	96.948	61.744	58.837
PUENTE ARANDA	803	4,50%	725.895	301.980	226.911	71.907	64.833
RAFAEL URIBE URIBE	605	3,39%	505.924	262.436	173.581	48.514	47.674
SAN CRISTÓBAL	766	4,29%	495.506	273.496	175.336	61.030	55.239
SANTA FE	808	4,53%	948.571	256.713	177.735	66.212	57.646
SUBA	1853	10,38%	1.104.388	315.501	255.685	89.827	76.632
SUMAPAZ	39	0,22%	486.213	487.949	44.139	78.179	70.769
TEUSAQUILLO	1314	7,36%	1.432.044	345.056	265.403	94.038	80.999
TUNJUELITO	474	2,66%	504.101	249.549	143.944	53.272	49.783
USAQUÉN	1294	7,25%	1.190.790	322.484	321.923	82.732	71.124
USME	592	3,32%	365.552	235.064	108.655	44.080	44.456

Tabla 3: Tabla de medias por localidad. Elaboración propia

Mientras que Sumapaz presenta el valor más bajo, con una diferencia de casi $278.000 frente a Usaquén, con $44.139. Para las variables NHCMP7AA y NHCMP7BA que describen el valor mensual promedio gastado en productos destinados al aseo del hogar y

aseo personal, se muestra que para los dos casos, Teusaquillo tiene el valor más elevado con $94.038 y $80.999 respectivamente. Mientras que Usme, para las dos variables mencionadas, también presenta los valores más bajos de toda la lista, con un valor cercano a los $44.000, siendo casi la mitad que el valor de la localidad de Teusaquillo.

	NHCCP10	MERCADO	SERVICIOS_PUB	NHCMP7AA	NHCMP7BA
count	17.849	17.849	17.849	17.849	17.849
mean	845.754	297.200	211.317	71.482	64.533
std	845.857	276.814	169.928	69.858	56.359
min	104.000	26.000	13.000	7.000	11.000
25%	450.000	150.000	110.000	30.000	30.000
50%	650.000	250.000	179.000	50.000	50.000
75%	1.000.000	380.000	264.728	100.000	80.000
max	20.000.000	8.150.000	3.488.000	1.500.000	1.000.000

Tabla 4: Estadísticos descriptivos. Elaboración propia.

Dentro de la muestra a trabajar, se evidencia que la variable que indica el valor pagado mensual por concepto de arriendo NHCCP10, tiene una promedio de $845.753, no obstante se tiene que alcanza un valor máximo de $20.000.000 que se puede considerar como un dato outlier, dado que la distribución entre el primer cuartil y el tercer cuartil oscila entre los $450.000 y $1.000.000.

Por su parte, la variable MERCADO, que indica el valor mensual que gasta el hogar para la compra de alimentos, la cual muestra que su valor mediano es de $250.000, mientras que alcanza un máximo de $8.150.000. Los servicios públicos, reflejan el valor mensual pagado por los hogares por este concepto para agua, acueducto, alcantarillados luz, internet y gas, donde tiene que en total hay valores entre $13.000 y $3.488.000

Por su parte, las variables NHCMP7AA y NHCMP7BA que indican cuánto destinan sus hogares de forma mensual a cubrir los gastos de aseo personal y aseo del hogar, tienen un comportamiento similar, dado que los valores entre el primer y tercer cuartil oscilan entre 30.000 y 100.000 aproximadamente, el promedio de la primera variable es de $71.481 y de la segunda variable es de $64.532

9. MODELADO

Las variables a imputar se clasifican en 4 grupos de acuerdo al porcentaje de valores ausentes que presente cada una, siendo del 2%, 5%, 10% y 20%.

9.1. Imputación por mediana

Para cada conjunto de datos previamente formado, se calcula de manera independiente el valor de la mediana para cada variable, el cual será reemplazado en cada uno de los faltantes de la variable correspondiente.

	2%	**5%**	**10%**	**20%**
NHCCP10	600.000	600.000	600.000	600.000
MERCADO	200.000	200.000	200.000	200.000
SERVICIOS_PUB	170.550	171.000	171.000	170.600
NHCMP7AA	50.000	50.000	50.000	50.000
NHCMP7BA	40.000	40.000	40.000	40.000

Tabla 5: Valor mediana. Elaboración propia

Los valores de la mediana calculados para variable de cada grupo de datos se muestra en la tabla anterior, donde no se evidencian grandes diferencias entre los valores calculados entre los diferentes niveles de porcentajes.

9.2. Imputación MICE

Con el paquete de R que lleva este mismo nombre, se desarrolla la imputación mediante el método “pmm”, cuyas siglas en inglés traducen “coincidencia de medias predictiva” en el cual se decide realizar 5 imputaciones múltiples, con una semilla de 500. Con la misma lógica del método anterior, se imputa a cada grupo de variables según su porcentaje de valores ausentes, obteniendo por cada uno 5 imputaciones diferentes de acuerdo al número de iteraciones seleccionadas.

Con el fin de unificar los valores plausibles obtenidos, son promediados para así calcular un valor único imputado por cada variable a estimar.

9.3. Imputación Miceforest

Al igual que la imputación por MICE, se realiza la imputación de cada conjunto de datos que previamente está clasificado por su nivel de ausentes. A través de la función MultipleImputedKernel de Python, se generan 5 datasets los cuales generan por cada iteración valores plausibles diferentes. Estos valores plausibles obtenidos, son promediados entre sí para obtener un único valor para cada campo faltante.

9.4. Imputación Red Neuronal

Para cada uno de los niveles de ausentes seleccionados, se entrena una red con la función MIDAS de python. Cada red alcanza su valor óptimo cuando tiene una layer structure de [256, 256] y un total de 70 epochs. Luego son generadas 35 muestras sobre la imputación entrenada, las cuales dan como resultado valores plausibles que son promediados para así tener un único valor por hogar para cada variables estimada.

De igual forma, la imputación propuesta con redes neuronales para data con faltantes al 20%, muestra en las estadísticas que a medida que aumenta el nivel de ausente en la data, mayor es la variación respecto a los datos originales, no obstante, se observa que con esté método, se evidencia una mejora respecto a MICEFOREST, ya que ahora la desviación estándar marcada por la variable NHCCP10 es de $765.093 mientras que la anterior fue de $753.363. Para todas las demás variables, los estadísticos resultantes son también mejores que los obtenidos a través de MICEFOREST.

10. RESULTADOS

10.1. Resultados error cuadrático medio

Para medir y comparar los resultados de cada método implementado, se utiliza el error cuadrático medio (RMSE) en cada conjunto de datos. Este mide la cantidad de error que hay entre dos conjuntos de datos, es decir, compara un valor predicho y un valor observado o conocido.

Para cada uno de métodos implementados se tienen los siguientes resultados:

10.1.1. Mediana

MÉTODO	VARIABLES	2%	5%	10%	20%
MEDIANA	NHCCP10	86.189	237.477	334.025	403.525
	MERCADO	29.187	51.800	94.174	124.413
	SERVICIOS_PUB	24.008	40.343	50.962	76.235
	NHCMP7AA	9.657	17.414	22.915	31.532
	NHCMP7BA	7.781	12.851	18.602	25.867

Tabla 6: RMSE para la mediana. Elaboración propia.

En la imputación por mediana, se evidencia que para todas las variables, el error aumenta en cada nivel de porcentaje establecido, siendo la variable valor pagado por concepto de arriendo (NHCCP10) la que mayor valor tiene para todos los casos. La variable NHCMP7BA es la que menor valor de RMSE tiene frente a las demás variables, pese a que su descripción y estadísticas guardan similitud con la variable NHCMP7AA.

10.1.2. MICE

MÉTODO	VARIABLES	2%	5%	10%	20%
MICE	NHCCP10	87.465	238.271	330.712	387.569
	MERCADO	29.070	50.484	92.814	127.404
	SERVICIOS_PUB	22.084	37.876	48.135	74.859
	NHCMP7AA	8.898	15.317	19.511	27.480
	NHCMP7BA	7.179	9.989	16.367	22.599

Tabla 7: RMSE para la MICE. Elaboración propia.

Al igual que la imputación por mediana, el error aumenta en la medida que aumenta el grado de valores perdidos, sin embargo, haciendo una comparación con el método anterior, se observa que a través de MICE hay mejora en la mayoría de las variables, siendo

MERCADO al 20% y NHCCP10 al 2% y 5% mejores con el método de la mediana, confirmando así las técnicas multivariadas presentan mejores resultados que las univariadas.

10.1.3. MICEFOREST

MÉTODO	VARIABLES	2%	5%	10%	20%
MICEFOREST	NHCCP10	81.158	232.617	326.400	390.702
	MERCADO	28.318	50.314	92.639	122.338
	SERVICIOS_PUB	22.618	38.692	48.748	73.442
	NHCMP7AA	9.088	16.344	21.216	29.477
	NHCMP7BA	7.150	12.091	17.673	24.322

Tabla 8: RMSE para la MICEFOREST. Elaboración propia.

Al igual que los casos anteriores, se evidencia que a mayor porcentaje de valores faltantes en la data, aumenta el RMSE, sin embargo, pese que tiene resultados similares a los obtenidos con MICE, para algunas variables y casos se ven mejoras con el método anterior. Por ejemplo, NHCMP7AA tiene por completo mejores resultados con MICE, SERVICIOS_PUB y NHCMP7BA, excepto cuando su nivel de faltantes es al 20% y 2% respectivamente, muestran mejores RMSE con el método anterior. En contraste, la variable mercado registra un mejor comportamiento con MICEFOREST, así como NHCCP10 cuando esta tiene porcentaje de faltantes diferentes al 20%.

10.1.4. Redes neuronales

MÉTODO	VARIABLES	2%	5%	10%	20%
REDES	NHCCP10	69.021	222.669	308.314	369.578
	MERCADO	28.042	49.090	91.305	121.594
	SERVICIOS_PUB	18.294	31.165	41.564	62.124
	NHCMP7AA	8.537	15.894	20.602	28.758
	NHCMP7BA	6.888	11.727	17.248	24.760

Tabla 9: RMSE para la red neuronal. Elaboración propia.

Finalmente, el método de redes muestra mejoras en la mayoría de las variables respecto a los demás métodos propuestos, variables como NHCCP10, MERCADO y SERVICIOS_PUB para todos los porcentajes de faltantes tiene el mejor RMSE con redes neuronales . Por su parte, las variables NHCMP7AA y NHCMP7BA para los niveles de ausentes del 5% al 20% presentan mejores resultados con la mediana y para los niveles del 2% mejor comportamiento con la redes.

10.2. Resultados distribuciones

Con el fin de validar el ajuste de las variables imputadas respecto a la data original que cuenta con todas las observaciones, se aplica el test de Kolmogorov-Smirnov con dos muestras a nuestro método objetivo, redes neuronales. Este test considera las siguientes hipótesis:

$$H_0: f(x) = f(x)'$$

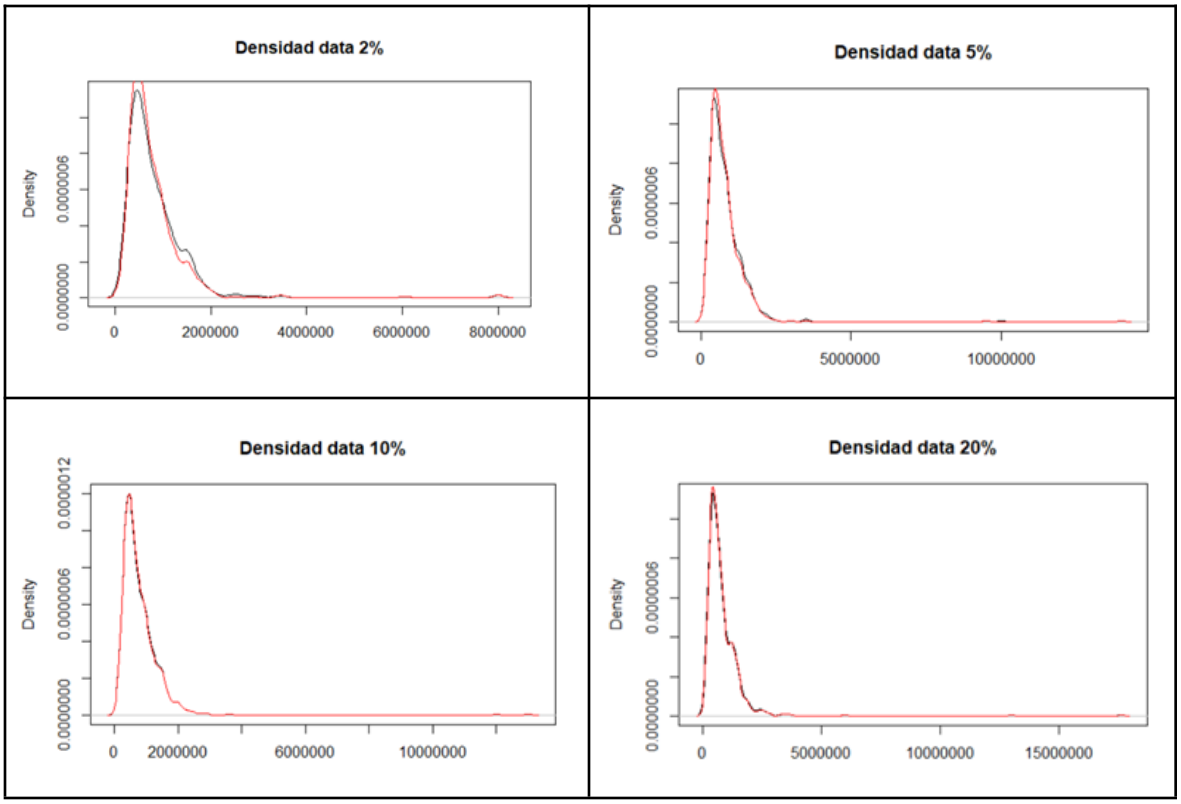
$$H_1: f(x) \neq f(x)'$$

Indicando si las dos distribuciones son iguales, se tiene los siguientes resultados para cada conjunto de datos con diferentes niveles de ausentes:

10.2.1. NHCCP10

%	D	p-value
Data 2%	0,054	0,4595
Data 5%	0,024	0,9987
Data 10%	0,006	0,999
Data 20%	0,01	0,999

Tabla 10: Prueba KS para NHCCP10 . Elaboración propia.

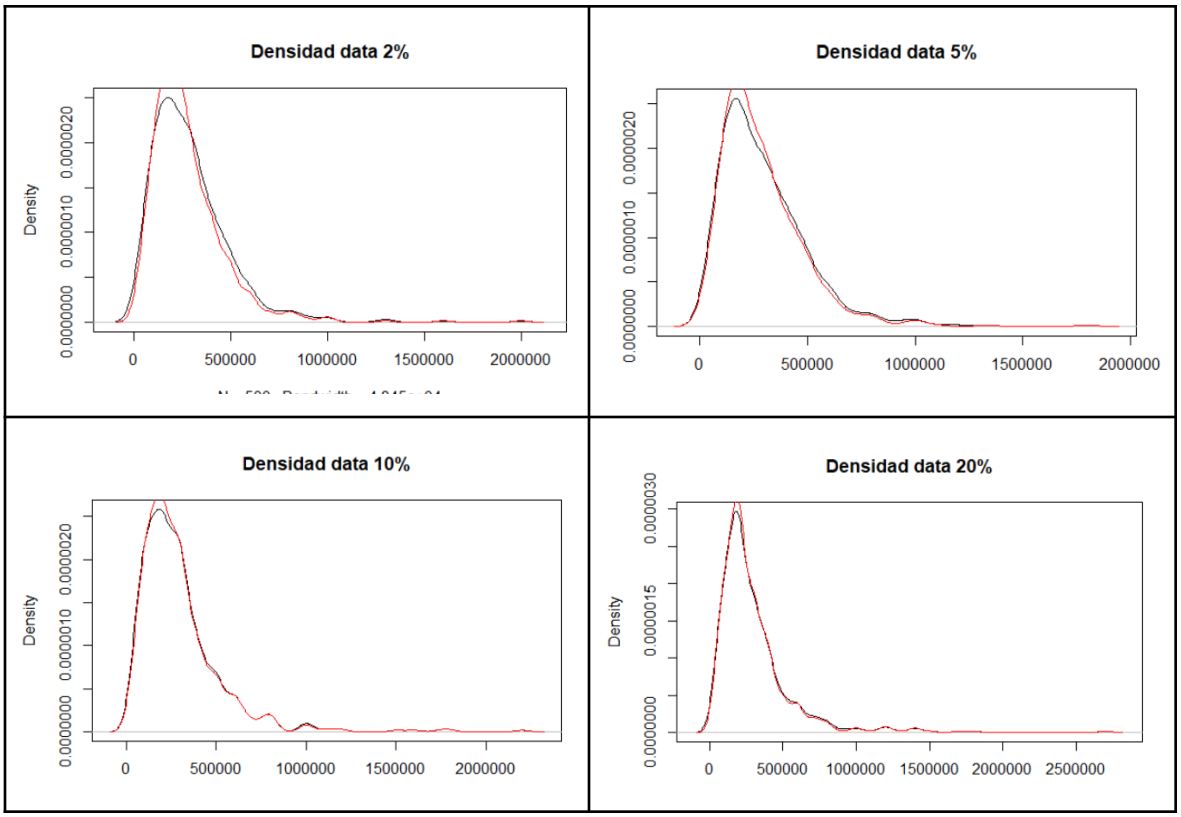


Esquema 4. Distribución original vs. estimada NHCCP10 . Elaboración propia.

10.2.2. MERCADO

%	D	p-value
Data 2%	0.076	0.1114
Data 5%	0.038	0.8632
Data 10%	0.014	0,999
Data 20%	0,01	0,999

Tabla 11 : Prueba KS para MERCADO. Elaboración propia.

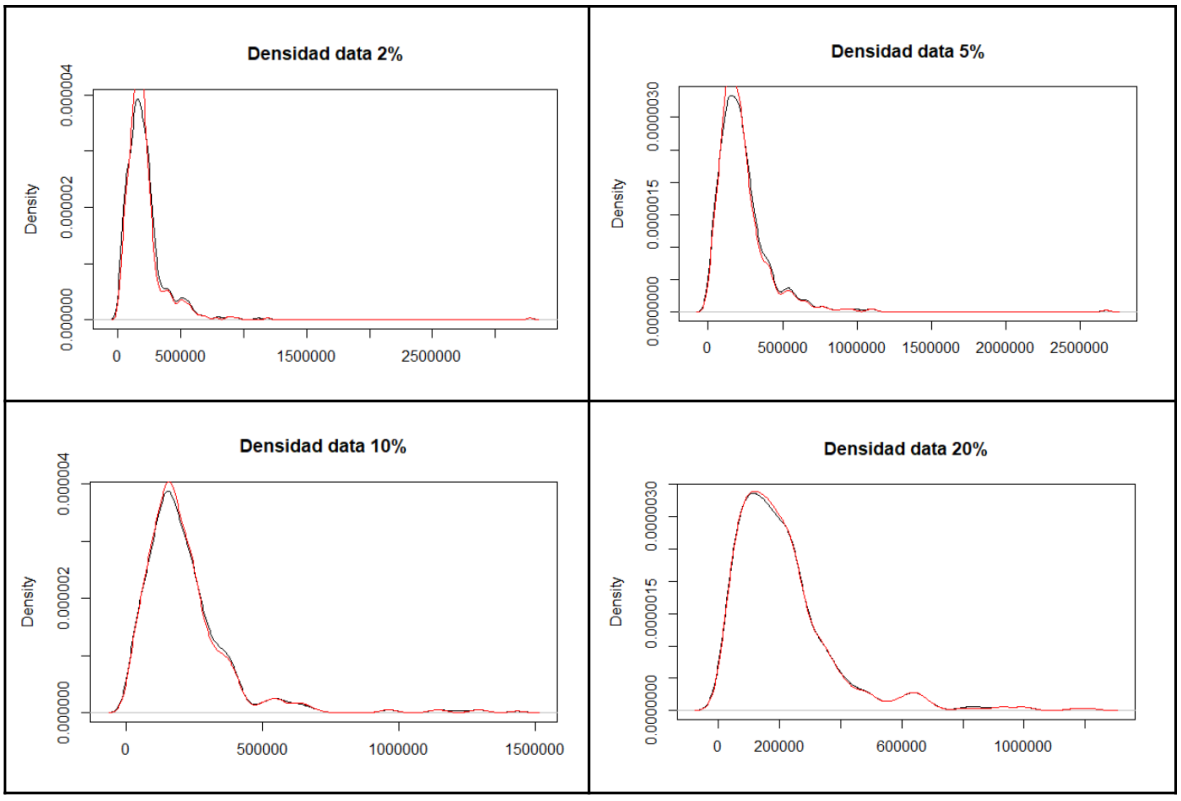


Esquema 5. Distribución original vs. estimada MERCADO. Elaboración propia.

10.2.3. SERVICIOS_PUB

%	D	p-value
Data 2%	0.076	0.1114
Data 5%	0.038	0.8632
Data 10%	0.016	0,999
Data 20%	0.006	0,999

Tabla 12 : Prueba KS para SERVICIOS_PUB. Elaboración propia.

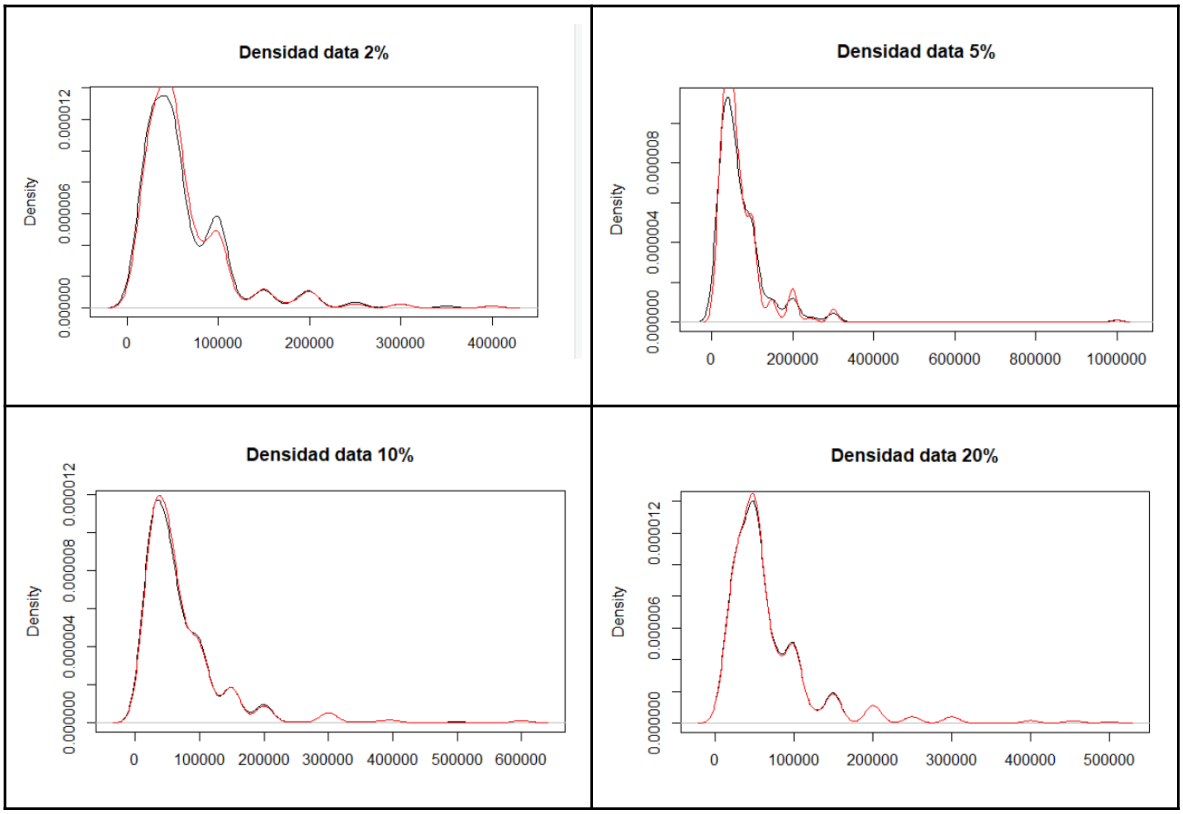


Esquema 6. Distribución original vs. estimada SERVICIOS_PUB. Elaboración propia.

10.2.4. NHCMP7AA

%	D	p-value
Data 2%	0.062	0.2917
Data 5%	0.034	0.9347
Data 10%	0.014	0.9999
Data 20%	0.008	0.9999

Tabla 13 : Prueba KS para NHCMP7AA. Elaboración propia.

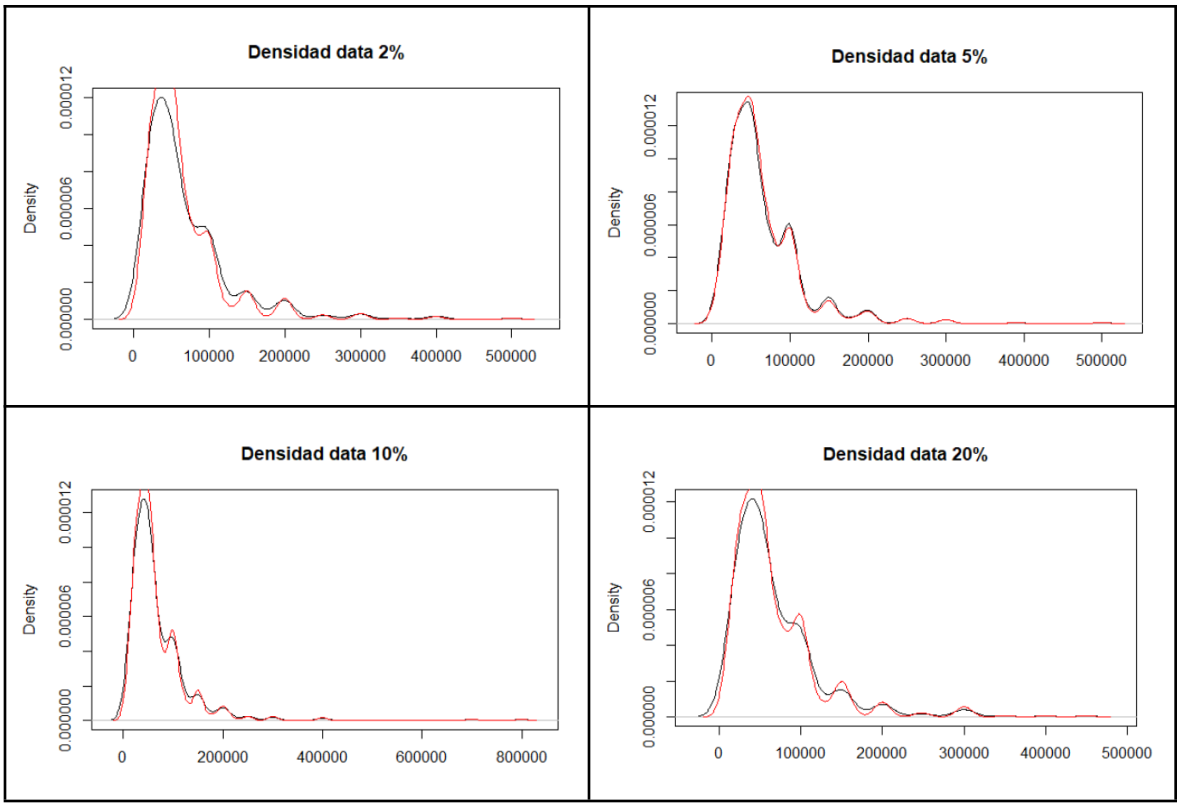


Esquema 7. Distribución original vs. estimada NHCMP7AA. Elaboración propia.

10.2.5. NHCMP7BA

%	D	p-value
Data 2%	0.064	0.2574
Data 5%	0.02	0.999
Data 10%	0.014	0.999
Data 20%	0.012	0.999

Tabla 14 : Prueba KS para NHCMP7BA. Elaboración propia.



Esquema 8. Distribución original vs. estimada NHCMP7AA. Elaboración propia.

11. CONCLUSIONES

En términos de métricas, tomando como referencia el RMSE, se encuentra que para las variables NHCCP10, MERCADO y SERVICIOS_PUB y las variables NHCMP7AA y NHCMP7BA cuando están con valores ausentes al 2% obtienen mejores resultados a través de redes neuronales, por su parte, las variables NHCMP7AA y NHCMP7BA cuando están con valores ausentes superiores 2%, presentan un mejor comportamiento tras su estimación con la MICE, es decir, que un 70% de la estimaciones son más eficientes utilizando redes.

Dado que es muy común encontrar valores fuera de rango o del contexto del objetivo de análisis, se lleva a cabo un proceso de transformación o limpieza, que consiste en tener en cuenta los campos que estuvieran por encima de determinado percentil, tal como se menciona en el apartado 7.2.2. transformación. Este proceso permitió mejorar las estimaciones de manera considerable, ya que sin transformación, se obtuvieron estimaciones con valores negativos, los cuales no presentan sentido en el contexto de los datos.

Dentro de las validaciones realizadas a los resultados de imputación a través de los diferentes métodos y porcentajes de valores ausentes, se encuentra que el modelo que utiliza redes neuronales, presenta la misma distribución que la data original, evidenciando así la eficiencia del método.

Adicional, es importante mencionar que si bien no todos los conjuntos de datos son MCAR, no hay variación o discriminación en los resultados con estos datos que cuenta con esta característica, dado a medida que disminuye el porcentaje de ausentes, mejores estimación se obtienen en cada método, asímismo, se logra identificar que independientemente del método que se utilice, ante el aumento de la tasa de faltantes en un conjunto de datos, la eficiencia disminuye frente a conjuntos con mayor número de campos observados.

12. REFERENCIAS

- Little, R. and Rubin, D., 2002. *Statistical analysis with missing data*. Hoboken: Wiley.
- Basogain, X., n.d. [online] Ocw.ehu.eus. Available at: <https://ocw.ehu.eus/pluginfile.php/40137/mod_resource/content/1/redes_neuro/contenidos/pdf/libro-del-curso.pdf>.

- Useche Castro, Lelly María, Mesa Ávila, Dulce Maria Una introducción a la Imputación de Valores Perdidos. Terra. Nueva Etapa [en línea]. 2006, XXII(31), 127-151 ISSN: 1012-7089. Disponible en: https://www.redalyc.org/articulo.oa?id=72103106.

- Li, C., 2013. Little's Test of Missing Completely at Random. *The Stata Journal: Promoting communications on statistics and Stata*, 13(4), pp.795-809.

- Azur, M., Stuart, E., Frangakis, C. and Leaf, P., 2011. Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, 20(1), pp.40-49.
- GitHub. 2021. *GitHub - AnotherSamWilson/miceforest: Multiple Imputation with Random Forests in Python*. [online] Available at: <https://github.com/AnotherSamWilson/miceforest> [Accessed 28 December 2021].
- Lall, R. and Robinson, T., 2021. The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Analysis*, pp.1-18.

- Javadi, S., Bahrampour, A., Saber, M. M., Garrusi, B., & Baneshi, M. R. (2021). Evaluation of four multiple imputation methods for handling missing binary outcome data in the presence of an interaction between a dummy and a continuous variable. Journal of Probability and Statistics, 2021, 1–14. https://doi.org/10.1155/2021/6668822
- (S/f). Mcgill.ca. Recuperado (2021), de https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In The Elements of Statistical Learning (pp. 587–604). Springer New York.
- Woodford, C. (2011, March 5). How neural networks work - A simple introduction. Explain That Stuff. https://www.explainthatstuff.com/introduction-to-neural-networks.html

Anexo 1

GRUPO	VARIABLE	DESCRIPCIÓN
Arriendo	NHCCP1	Indica si la vivienda es en arriendo o no
	NHCCP10	Cuánto paga de arriendo o leasing
Características del hogar	NHCCPCTRL2	Cuántas personas componen el hogar
	NHCCP19	Cantidad de cuartos sin baños, cocinas, garajes
	NHCCP22A	Jardín o patio
	NHCCP22B	Lote o solar
	NHCCP22C	garaje
	NHCCP22D	Terraza
	NHCCP22E	Zonas verdes
	NHCCP22F	Zonas comunes
	NHCCP22G	Ninguna
	NHCCP24	Espacio exclusivo para preparar alimentos
	NHCCP25	Cocina de uso exclusivo para personas del hogar
	NHCCP26	Energía o combustible para cocinar
	NHCCP29	Servicio de agua 24 hrs
	NHCCP31	Tipo de sanitario
	NHCCP36A	Lavamanos
	NHCCP36B	Lavadero
	NHCCP36C	Tanque de reserva de agua
	NHCCP36D	Ninguna
Servicios públicos y de TI	NHCDP1	¿Paga acueducto?
	NHCDP2	Valor acueducto
	NHCDP3	¿ Paga alcantarillado?
	NHCDP4	Valor alcantarillado
	NHCDP5	¿Paga recolección de basuras?
	NHCDP6	Valor recolección de basuras
	NHCDP9	¿Paga energía eléctrica?

	NHCDP11	Valor energía eléctrica
	NHCDP16	¿Paga gas natural?
	NHCDP18	Valor gas natural
	NHCDP28	¿Paga internet?
	NHCDP29	Valor internet
	SERVICIOS_PUB	Valor total de servicios pagados en el hogar
Características de la vivienda	NVCBP1	Vía de acceso a la edificación
	NVCBP2	Estado de la vía
	NVCBP3	Tiene andén
	NVCBP4	Está en conjunto cerrado
	NVCBP5	Iluminación suficiente
	NVCBP6	Número de pisos de la edificación
	NVCBP7	Tiene ascensor
	NVCBP10C	Tipo casa
	NVCBP10A	Tipo apartamento
	NVCBP10H	Tipo cuarto
	NVCBP10O	Tipo otro
	NVCBP11A	Energía eléctrica
	NVCBP11B	Acueducto
	NVCBP11C	Alcantarillado
Problemas de la vivienda	NVCBP8A	El hogar no cuenta con problemas de humedad en techo y paredes
	NVCBP8B	El hogar no cuenta con problemas de goteras en el techo
	NVCBP8C	El hogar no cuenta con problemas de grietas en techos
	NVCBP8D	El hogar no cuenta con problemas de fallas en tuberías
	NVCBP8E	El hogar no cuenta con problemas de grietas en el piso
	NVCBP8F	El hogar no cuenta con problemas de tejas en mal

		estado
	NVCBP8G	El hogar no cuenta con problemas de poca ventilación
	NVCBP8H	El hogar no cuenta con problemas de inundación por lluvias
	NVCBP8I	El hogar no cuenta con problemas de peligro de derrumbe y avalancha
	NVCBP8J	El hogar no cuenta con problemas de hundimiento de terreno
	NVCBP11D	El hogar no cuenta con problemas de recolección de basuras
	NVCBP14A	El hogar no cuenta con problemas de cerca de fábricas
	NVCBP14B	El hogar no cuenta con problemas de cerca de basureros
	NVCBP14C	El hogar no cuenta con problemas de cerca de plazas de mercado o mataderos
	NVCBP14D	El hogar no cuenta con problemas de cerca de terminales de buses
	NVCBP14E	El hogar no cuenta con problemas de cerca a bares o prostíbulos
	NVCBP14F	El hogar no cuenta con problemas de cerca de expendio de drogas
	NVCBP14G	El hogar no cuenta con problemas de cerca de terrenos baldíos
	NVCBP14H	El hogar no cuenta con problemas de cerca a líneas de alta tensión
	NVCBP14I	El hogar no cuenta con problemas de cerca de caños de aguas residuales
	NVCBP14J	El hogar no cuenta con problemas de cerca a zonas de riesgo de incendio
	NVCBP14K	El hogar no cuenta con problemas de cerca a talleres y gasolineras

	NVCBP15A	El hogar no cuenta con problemas de ruido
	NVCBP15B	El hogar no cuenta con problemas de exceso de anuncios publicitarios
	NVCBP15C	El hogar no cuenta con problemas de inseguridad
	NVCBP15D	El hogar no cuenta con problemas de contaminación del aire
	NVCBP15E	El hogar no cuenta con problemas de malos olores
	NVCBP15F	El hogar no cuenta con problemas de manejos inadecuado de basuras
	NVCBP15G	El hogar no cuenta con problemas de invasión espacio público
	NVCBP15H	El hogar no cuenta con problemas de animales que molesten
Ocupación	NPCKP1T	Trabajó
	NPCKP1B	Buscó trabajo
	NPCKP1E	Estudió
	NPCKP1H	Realizó labores de hogar
	NPCKP1I	Incapacitado parcial o permanente
	NPCKP1O	Se ocupó en otra cosa
Subsidios	NPCKP29	Recibe subsidio alimentación
	NPCKP29A	Valor subsidio alimentación
	NPCKP30	Auxilio de transporte
	NPCKP30A	Valor auxilio de transporte
	NPCKP31	Subsidio familiar
	NPCKP31A	Valor subsidio familiar
	NPCKP32	Recibe subsidio educativo
	NPCKP32A	Valor subsidio educativo
	SUBSIDIOS	Monto total de subsidios por hogar
Ingresos	NPCKP52	Paga pensión
	NPCKP52A	Valor pensión
	NPCKP23	Monto ingreso por empleo

Gastos	MERCADO	Monto total gastado el alimentos
	NHCMP5A	¿Alcohol y cigarros en los últimos 7 días?
	NHCMP5AA	Alcohol y cigarro
	NHCMP5B	?¿Transporte en los últimos 7 días
	NHCMP5BA	Transporte
	NHCMP5D	¿Combustible y parqueadero en los últimos 7 días?
	NHCMP5DA	Combustible y parqueadero
	NHCMP5E	Comida fuera de casa
	NHCMP5EA	¿Comida fuera de casa en los últimos 7 días?
	NHCMP5F	¿Apuestas en los últimos 7 días?
	NHCMP5FA	Apuestas
	NHCMP7A	¿Productos de aseo casa último mes?
	NHCMP7AA	Productos aseo casa
	NHCMP7B	¿Aseo personal último mes?
	NHCMP7BA	Aseo personal
	NHCMP7G	¿Entretenimiento último mes?
	NHCMP7GA	Entretenimiento