Revista esprint Investigación

https://rei.esprint.tech

ISSN: 2960-8317

Vol. 4 N° 2, julio-diciembre 2025 (358-375)

Artículo de investigación

Random Forest como herramienta para mejorar la precisión en la imputación de datos meteorológicos en Chimborazo, Ecuador

Random Forest as a tool to improve accuracy in the imputation of meteorological data in Chimborazo, Ecuador

Laura Susana Naranjo Ordoñez*
Universidad Politécnica Estatal del Carchi
Tulcán - Ecuador
Instituto Superior Universitario "San Gabriel"
Riobamba - Ecuador
laura.naranjo@upec.edu.ec
https://orcid.org/0000-0002-3534-0545

Amalia Isabel Escudero Villa
Universidad Politécnica Estatal del Carchi
Tulcán - Ecuador
Escuela Superior Politécnica de Chimborazo
Riobamba - Ecuador
aescudero@espoch.edu.ec
https://orcid.org/0000-0003-4506-932X

***Correspondencia:**
laura.naranjo@upec.edu.ec

Cómo citar este artículo:
Naranjo, L., & Escudero, A. (2025). Random Forest como herramienta para mejorar la precisión en la imputación de datos meteorológicos en Chimborazo, Ecuador. *Esprint Investigación, 4*(2), 358-375. https://doi.org/10.61347/ei.v4i2.169

Recibido: *23 de agosto de 2025*
Aceptado: *23 de septiembre de 2025*
Publicado: *30 de septiembre de 2025*

Resumen: Gestionar la información faltante en los registros meteorológicos monitoreados por el Grupo de Energías Alternativas y Ambientales (GEAA) es esencial para realizar un análisis climático preciso y tomar decisiones informadas. El objetivo de este artículo fue evaluar la efectividad del algoritmo Random Forest mediante el software estadístico R. La investigación tuvo un enfoque cuantitativo con un alcance descriptivo-comparativo; el diseño es no experimental y longitudinal. Se comparó Random Forest (k-NN) con el de k-vecinos más cercanos (k-NN o K-Nearest Neighbors) utilizando diversas métricas, como Error Cuadrático Medio (RMSE), Error Medio Absoluto (MAE), pruebas de Kolmogorow-Smirnov, sumado a ello la eficiencia computacional en cuanto al tiempo y memoria. Los resultados indicaron que Random Forest obtuvo mayor precisión con respecto a k-NN; los valores de RMSE y MAE son evidentemente más bajos. RF demandó mayor recurso computacional, la capacidad y efectividad al momento de procesar registros de alta complejidad lo convierten en la mejor opción, proporcionando mayor confiabilidad al momento de imputar y, por ende, datos de calidad.

Palabras clave: Datos faltantes, imputación, k-vecinos más cercanos (k-NN), meteorología, Random Forest (RF).

Abstract: *Managing missing information in meteorological records monitored by the Alternative and Environmental Energy Group (GEAA) is essential for conducting accurate climate analyses and making informed decisions. The objective of this article was to evaluate the effectiveness of the Random Forest algorithm using the statistical software R. The research followed a quantitative approach with a descriptive-comparative scope; the design was non-experimental and longitudinal. Random Forest was compared with the k-Nearest Neighbors (k-NN) algorithm using various metrics, such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Kolmogorov-Smirnov tests, along with computational efficiency in terms of time and memory. The results indicated that Random Forest achieved higher accuracy compared to k-NN, with significantly lower RMSE and MAE values. Although RF required greater computational resources, its capacity and effectiveness in processing highly complex records make it the best option, providing greater reliability in data imputation and, consequently, higher-quality datasets.*

Keywords: *Imputation, k-nearest neighbors (k-NN), meteorology, missing data, Random Forest (RF).*

1. Introducción

Las estaciones meteorológicas generan registros que permiten transformar la información cuantificable en insumos para la toma de decisiones. Para que estos datos adquieran valor científico, deben pasar por un proceso de validación que garantice bases confiables, reduzca los márgenes de error y aumente la precisión de los resultados. Sin embargo, en el caso de estudio compuesto por 11 estaciones meteorológicas en la provincia de Chimborazo-Ecuador, se evidencian problemas recurrentes en los registros debido al alto porcentaje de vacíos. Esta situación limita la posibilidad de realizar predicciones confiables, introduce sesgos en los análisis y dificulta la toma de decisiones (Troyanskaya et al., 2001).

En meteorología, el desafío se intensifica por la cantidad de variables a considerar. La temperatura es una de las más analizadas (Economou et al., 2023), pero también influyen factores como la humedad, la presión atmosférica, la precipitación, el viento, la nubosidad y la radiación solar (Zúñiga & Crespo, 2021). Estos parámetros, al estar interrelacionados, requieren registros completos para reflejar con fidelidad las condiciones atmosféricas de un lugar.

La obtención de tales registros depende de diversos métodos y tecnologías de monitoreo que capturan la dinámica del clima (Dimitri et al., 2024). No obstante, persiste un problema común: la presencia de datos incompletos, producto de fallos en sensores, interrupciones en la transmisión o errores humanos. Estas deficiencias reducen la disponibilidad de información climática integral y, por ende, limitan la calidad de los análisis (Hayawi et al., 2024).

En respuesta a este reto, en los últimos años se ha incorporado tecnología como el Internet of Things (IoT), que ha mejorado los procesos de recopilación (Saied & Guirguis, 2025). Pese a su utilidad, la implementación de estas herramientas aún enfrenta restricciones técnicas y operativas (Galván & Medina, 2007; Breiman, 2001). Como resultado, el análisis meteorológico puede verse retrasado y, en algunos casos, derivar en pronósticos erróneos (Lacourly, 2012). Por ello, disponer de registros completos y confiables se vuelve una condición indispensable, más aún cuando continuamente se desarrollan modelos predictivos que buscan optimizar la toma de decisiones (Rizal et al., 2022).

En este contexto, la gestión de los datos adquiere un papel central. Entre las estrategias más utilizadas para imputar vacíos se encuentra el algoritmo k-Nearest Neighbors (k-NN), que funciona como un método de aprendizaje supervisado no paramétrico. Este evalúa la proximidad de k vecinos definidos por el usuario, aplicando fórmulas como las distancias euclidianas y de Manhattan. No obstante, el rendimiento del modelo depende en gran medida de la correcta elección de k, lo que puede comprometer los resultados (Hou et al., 2024; Troyanskaya et al., 2001).

Con el fin de superar estas limitaciones y garantizar la calidad de los registros meteorológicos de Chimborazo-Ecuador, monitoreados por el GEAA, esta investigación se propuso evaluar la efectividad del algoritmo Random Forest mediante el software estadístico R. A diferencia de k-NN, este método ofrece un modelo más robusto, capaz de aprovechar la correlación entre múltiples variables para predecir valores faltantes con menor sesgo. Su funcionamiento se basa en la construcción de árboles de decisión a partir de muestras bootstrap, lo que introduce diversidad y reduce el sobreajuste (Yarupaita, 2021; Han et al., 2020).

Cada árbol aporta un voto y la predicción final se obtiene mediante agregación, un esquema que permite manejar grandes volúmenes de datos y que muestra resistencia frente al ruido y a los valores atípicos (Breiman, 2001; Schonlau & Zou, 2020; Bashir et al., 2024). Estas características convierten a Random Forest en una alternativa sólida para la imputación de datos meteorológicos, con potencial de ser replicada en otras regiones que enfrentan dificultades similares en la validación de registros (Díaz-Uriarte & Alvarez, 2006).

2. Metodología

La investigación tiene un enfoque cuantitativo con un alcance descriptivo-comparativo (Hernández-Sampieri & Mendoza, 2018). El diseño es no experimental y longitudinal, puesto que se analizaron la frecuencia y la distribución temporal de faltantes en las estaciones meteorológicas del GEAA en Chimborazo–Ecuador. Se identificaron los registros por periodos de tiempo en mes, día y hora, correspondientes al año 2019, con el fin de realizar la imputación y la evaluación de la efectividad de los métodos Random Forest y k-NN.

Random Forest (RF)

Se aplicó el algoritmo del paquete Random Forest en R, este utiliza 300 árboles y 10 iteraciones. El método permite observar relaciones no lineales entre las variables, es eficiente ante valores extremos. La elección se debió a la seguridad que tiene este cuando se trabajan múltiples variables correlacionadas (Amat, 2020; Hastie et al., 2009).

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

$\hat{f}(x) =$ estimación,

$B =$ número total de árboles,

$T_b(x) =$ predicción individual de b-ésima del árbol de decisiones.

k-vecinos más cercanos (k-NN o K-Nearest Neighbors)

Para la comparación se utilizó el algoritmo k-NN del paquete VIM con k=5. El método imputa datos con base en los valores más similares, evaluando la distancia euclidiana entre observaciones (Amat, 2020).

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Donde:

$d(x, y) =$ distancia euclídea,

$x, y =$ puntos del vector,

$y =$ segundo punto del vector,

$x_i, y_i \;\; =$ coordenadas i-ésima,

$n =$ número de dimensiones o atributos.

Evaluación

Para validar la calidad de la imputación, se trabajó con un conjunto de datos reales completos, en los cuales se colocaron valores faltantes de manera aleatoria para simular pérdidas. Para la comparación de las bases

imputadas y originales se utilizaron métricas como el Error Cuadrático Medio (RMSE), que mide el error cuadrático promedio, y el Error Absoluto Medio (MAE), que evalúa el promedio del error absoluto.

Se emplearon contrastes de hipótesis para evaluar parámetros necesarios, tales como: el test de Kolmogorov-Smirnov para comparar las distribuciones de las bases imputadas con las originales; la prueba de Shapiro-Wilk para identificar si los datos provienen de poblaciones con distribución normal; y las pruebas t de Student y Wilcoxon para comparar las medias en cuanto al tiempo y la capacidad de memoria requerida por cada técnica de imputación (Amat, 2020).

$$D = \sup_{1 \le i \le n} \left| \widehat{F}_n(x_i) - F_0(x_i) \right|$$

Donde:

$x_i = i$-ésimo valor observado en la muestra,

$\hat{F}_n(x_i) =$ estimador de la probabilidad,

$F_0(x_i) =$ probabilidad de observar valores menores o iguales que x_i.

Los resultados obtenidos fueron tabulados por variable, método, y el desempeño se visualizó mediante figuras y tablas para una mejor comprensión del comportamiento de cada imputador.

Recursos

Los recursos empleados para el desarrollo del estudio incluyeron el lenguaje R (versión ≥ 4.2.0), organizado por fases. Los resultados fueron exportados a hojas de Excel para facilitar su manejo. Se integraron librerías esenciales como openxlsx, lubridate, tidyverse, randomForest, missForest, mice, VIM, writexl, zoo, y dplyr.

Datos y Variables

El conjunto de datos empleado corresponde a los registros de las estaciones meteorológicas ubicadas en Urbina, Matus, San Juan, Espoch, Quimiag, Tunshi, Alao, Multitud, Tixán y Cumandá durante el año 2019, pertenecientes a la provincia de Chimborazo-Ecuador (figura 1).

Figura 1

Localización de las estaciones meteorológicas monitoreadas por el GEAA.

La base está compuesta por 20 variables, en su mayoría cuantitativas y continuas, medidas cada hora, lo que ofrece una visión detallada de las condiciones atmosféricas y del suelo. Entre ellas se encuentran la temperatura ambiental, la humedad relativa, la presión atmosférica, la radiación solar, la temperatura del suelo a distintos niveles, la precipitación, la dirección y velocidad del viento, y la sensación térmica, como se describe en la tabla 1.

Tabla 1

Variables meteorológicas analizadas en la investigación

Nº	Abreviatura	Significado	Tipo	Unidad
$\mathbf{X_1}$	Stat_TA_1h_Avg	Temperatura ambiental promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_2}$	Stat_RH_1h_Avg	Humedad relativa promedio (cada hora)	Cuantitativa continua	%
$\mathbf{X_3}$	Stat_PA_1h_Avg	Presión atmosférica promedio (cada hora)	Cuantitativa continua	hPa
$\mathbf{X_4}$	Stat_SR_Dif_1h_Avg	Radiación solar difusa promedio (cada hora)	Cuantitativa continua	W/m^2
$\mathbf{X_5}$	Stat_SR_Glob_1h_Avg	Radiación solar global promedio (cada hora)	Cuantitativa continua	W/m^2
$\mathbf{X_6}$	Stat_TS_1h_TG1_Avg	Temperatura del suelo nivel 1 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_7}$	Stat_TS_1h_TG2_Avg	Temperatura del suelo nivel 2 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_8}$	Stat_TS_1h_TG3_Avg	Temperatura del suelo nivel 3 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_9}$	Stat_TS_1h_TG4_Avg	Temperatura del suelo nivel 4 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_{10}}$	Stat_TS_1h_TG5_Avg	Temperatura del suelo nivel 5 Promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_{11}}$	Stat_TS_1h_TG6_Avg	Temperatura del suelo nivel 6 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_{12}}$	Stat_TS_1h_TG7_Avg	Temperatura del suelo nivel 7 promedio (cada hora)	Cuantitativa continua	°C
$\mathbf{X_{13}}$	Sum_PR_1h_Sum	Precipitación acumulada (cada hora)	Cuantitativa continua	mm
$\mathbf{X_{14}}$	WRun	Recorrido del viento	Cuantitativa continua	m
$\mathbf{X_{15}}$	DirAvg	Dirección del viento promedio (cada hora)	Cuantitativa continua	Grados (°)
$\mathbf{X_{16}}$	GustDir	Dirección de la racha de viento	Cuantitativa continua	Grados (°)

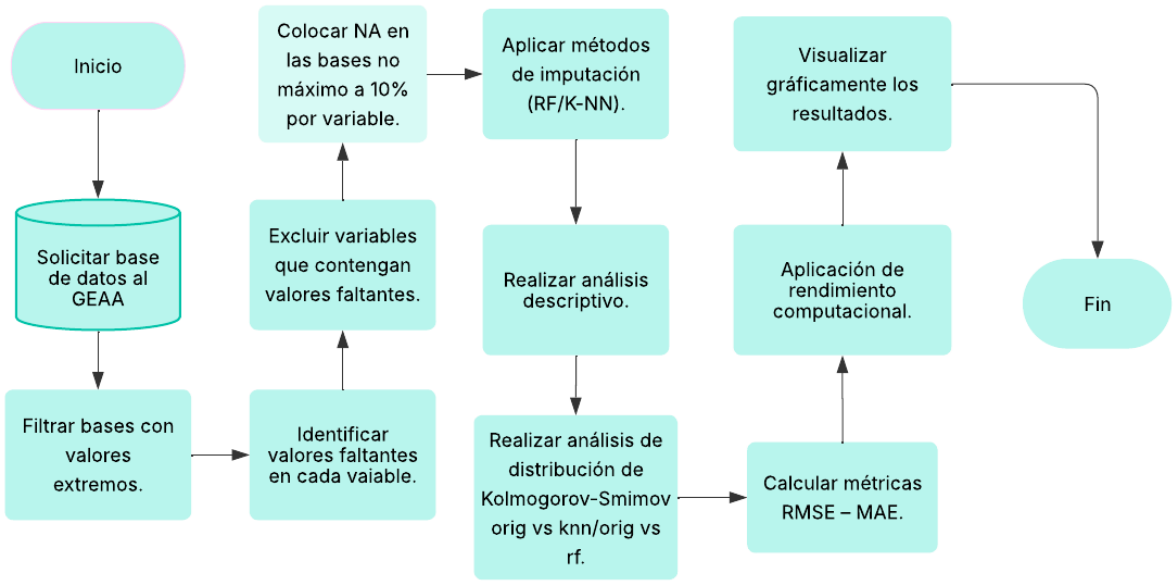
$\mathbf{X_{17}}$	GustH	Hora de la racha de viento	Cuantitativa continua	Hora (formato 24h)
$\mathbf{X_{18}}$	GustM	Minuto de la racha de viento	Cuantitativa continua	Minuto
$\mathbf{X_{19}}$	SpdAvg	Velocidad promedio del viento	Cuantitativa continua	m/s
$\mathbf{X_{20}}$	Stat_WindChill_1h_SDI12_Avg	Sensación térmica promedio (cada hora)	Cuantitativa continua	°C

Proceso de la data

El diagrama de flujo (figura 2) muestra paso a paso el proceso metodológico realizado para evaluar el mejor método de imputación de datos faltantes en cada variable. Se inicia con la recepción de la data y la verificación de registros mediante un análisis descriptivo, con el fin de comprobar que los datos cumplieran los parámetros establecidos por el GEAA, así como la omisión de variables con faltantes excesivos para lograr un análisis robusto. Posteriormente, se aplicaron las técnicas y métodos especificados en el diagrama.

Figura 2

Diagrama de flujo del proceso metodológico del estudio.



3. Resultados

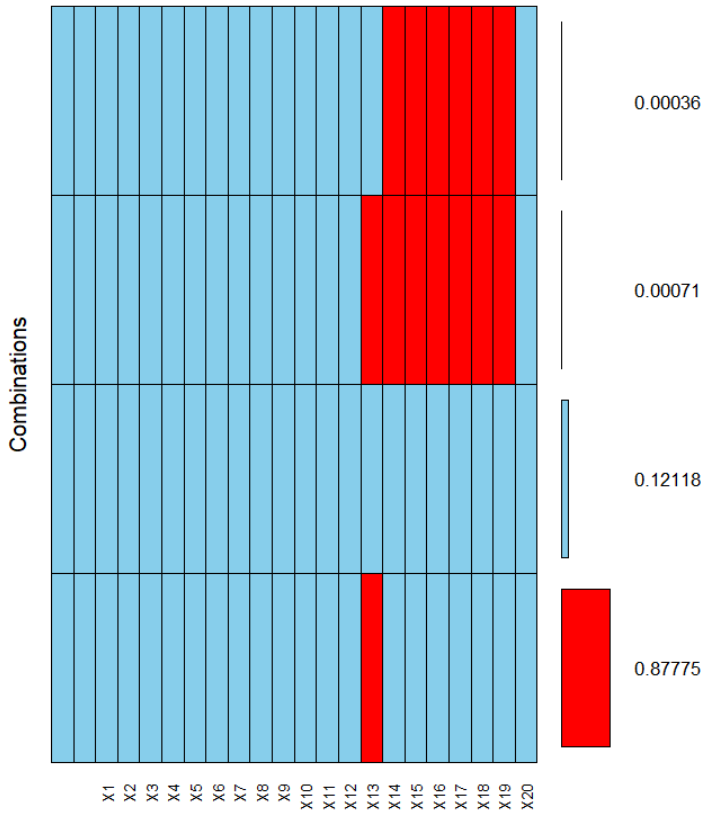
Esta sección presenta los hallazgos obtenidos al aplicar los métodos de imputación Random Forest y k-vecinos más cercanos, en los datos meteorológicos monitoreados por el GEAA en Chimborazo. Se evaluaron 20 variables (X_1,X_2,...,X_20) como indica la Tabla 1, de las cuales se excluyeron de la X_13 a la X_19 debido a la presencia de datos faltantes. Se aplicaron las métricas de desempeño: el MAE (error absoluto medio) y el RMSE (error cuadrático medio).

En la figura 3 se visualiza en el eje de las X cada una de las variables y el eje Y el porcentaje de faltantes. Las celdas azules representan los datos válidos, mientras que las celdas rojas indican ausencia de registros. Se observa que la mayoría de las variables están completas, a excepción de las X_13 a la X_19. Debido a que presentan porcentajes mayores al 10% de datos faltantes, con el propósito de obtener resultados más confiables para la investigación.

Dado que este estudio se centra en evaluar la efectividad de los métodos de imputación, era necesario asegurar que los resultados comparativos reflejaran la calidad del algoritmo y no las limitaciones de un conjunto de datos con pérdidas excesivas. Por esta razón, se priorizó trabajar únicamente con variables que presentaban registros prácticamente completos, lo que garantizó mayor confiabilidad en las conclusiones obtenidas.

Figura 3

Datos faltantes



Se visualizan los resultados del análisis descriptivo comparativo (tabla 2) de las bases imputadas y los datos originales, donde se observan los 281 valores faltantes creados aleatoriamente en todas las variables. Estos, una vez aplicados los métodos Random Forest y k-Nearest Neighbors, desaparecen. Se visualiza que las medidas de tendencia central (media, mediana y moda) y las medidas de dispersión (mínimo y máximo) se mantienen similares con respecto a la base original, lo que indica que la imputación no alteró los datos originales entregados por el GEAA en Chimborazo-Ecuador.

Tabla 2

Análisis estadístico descriptivo de los datos disponibles

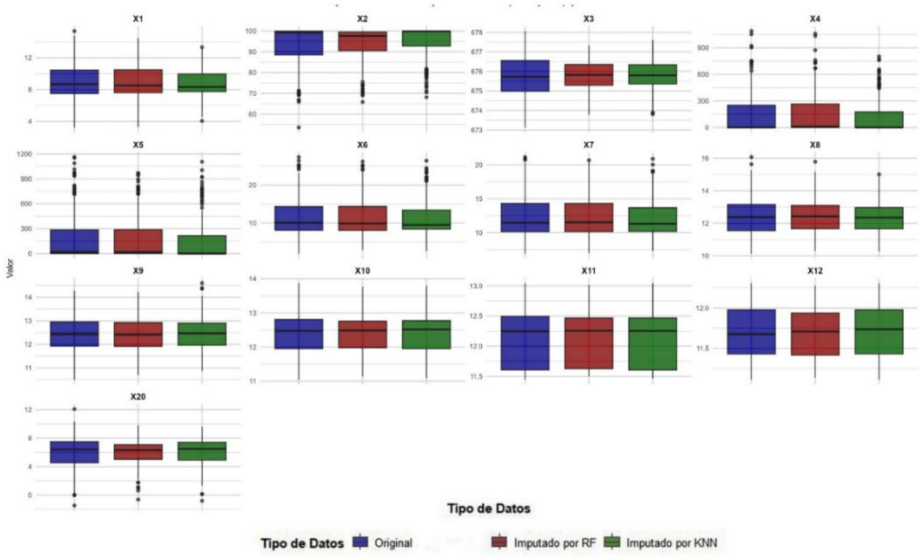
Variable	Base	NAs	Media	Mediana	Min	Max
	Original	281	11.171	10.491	0.521	20.920
$\mathbf{X_1}$	RF	0	11.139	10.477	0.521	20.92
	k-NN	0	11.157	10.503	0.521	20.92
	Original	281	79.651	83.961	33.247	99.006
$\mathbf{X_2}$	RF	0	79.924	84.347	33.247	99.006
	k-NN	0	79.751	83.933	33.247	99.006
	Original	281	709.038	709.174	704.241	713,031
$\mathbf{X_3}$	RF	0	709.001	709.125	704.241	713.031
	k-NN	0	709.008	709.136	704.241	713.031
	Original	281	134.011	0.002	0	1363.04
$\mathbf{X_4}$	RF	0	132.004	0	0	1363.04
	k-NN	0	135.436	0.217	0	1363.04
	Original	281	165.664	0	0	1383.878
$\mathbf{X_5}$	RF	0	163.794	0	0	1383.878
	k-NN	0	167.062	0.026	0	1383.878
	Original	281	13.072	10.781	-1.187	37.126
$\mathbf{X_6}$	RF	0	12.987	10.766	-1.187	37.126
	k-NN	0	13.042	10.806	-1.187	37.126
	Original	281	13.865	11.606	-0.283	39.970
$\mathbf{X_7}$	RF	0	13.779	11.582	-0.283	39.97
	k-NN	0	13.837	11.630	-0.283	39.97
	Original	281	14.758	13.572	3.491	33.078
$\mathbf{X_8}$	RF	0	14.746	13.612	3.491	33.078
	k-NN	0	14.766	13.630	3.491	33.078
	Original	281	14.964	14.435	6.251	27.723
$\mathbf{X_9}$	RF	0	15.009	14.497	6.251	27.723
	k-NN	0	14.982	14.469	6.251	27.723
	Original	281	14.940	14.933	9.438	20.505
$\mathbf{X_{10}}$	RF	0	15.014	15.039	9.438	20.505
	k-NN	0	14.955	14.948	9.438	20.505

	Original	281	14.792	14.939	11.693	17.155
$\mathbf{X_{11}}$	RF	0	14.836	15.086	11.693	17.155
	k-NN	0	14.793	14.935	11.693	17.155
	Original	281	14.464	14.510	12.388	15.961
$\mathbf{X_{12}}$	RF	0	14.495	14.702	12.388	15.961
	k-NN	0	14.446	14.394	12.388	15.961
	Original	281	5.976	6.346	-3.895	13.895
$\mathbf{X_{20}}$	RF	0	5.958	6.334	-3.895	13.895
	k-NN	0	5.970	6.362	-3.895	13.895

En la figura 4 se observa por cada uno de los elementos, un conjunto de tres diagramas de caja y bigotes, en los cuales se observa que las medidas de tendencia central y de dispersión, son similares con respecto a la base original ratificando que los métodos de imputación Random Forest y por k-vecinos más cercanos conservan la distribución original.

Figura 4

Distribución comparativa entre los datos originales, imputación con RF y k-NN.

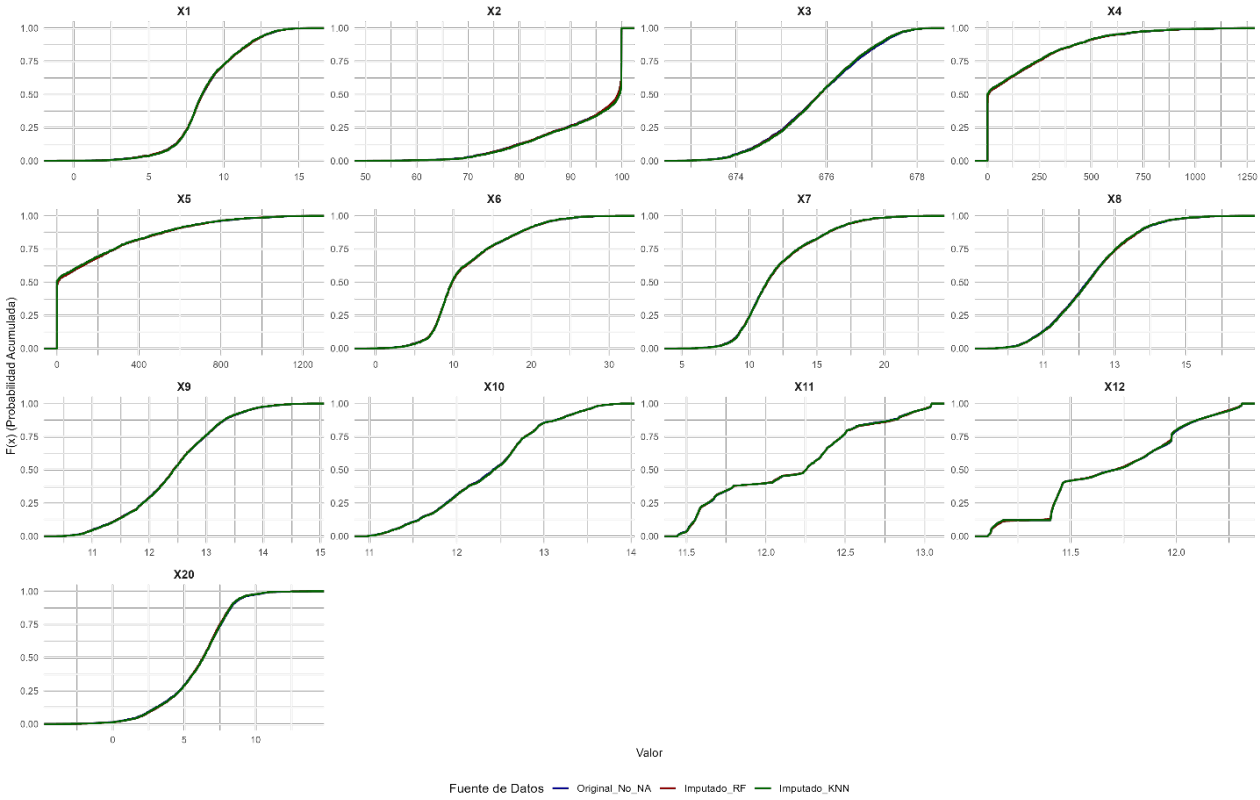


Análisis de distribución de Kolmogorov-Smirnov por variable y método de imputación

En la figura 5 se observan las Funciones de Distribución Acumulativa Empírica (FDCE) de cada variable; se representa gráficamente la probabilidad acumulativa, muestra la proporción de puntos de datos menores o iguales a un valor dado. Estos gráficos escalonados van de 0 a 1 en el eje Y; las líneas estrechamente alineadas sugieren distribuciones estadísticamente iguales, lo que proporciona una visión de los efectos de imputación.

Figura 5

Comparación de distribuciones de Kolmogorov-Smirnov por variable y método de imputación



Para determinar si el método de imputación de Random Forest supera significativamente al de los k-vecinos más cercanos, se evaluó la consistencia entre las distribuciones imputadas y el conjunto de datos original. La siguiente sección describe la formulación de la hipótesis.

$\boldsymbol{H_0}$: No existe una diferencia significativa entre la distribución de la data original con respecto a las imputadas mediante Random Forest y k-vecinos más cercanos.

$\boldsymbol{H_1}$: Existe una diferencia significativa entre la distribución de la data original con respecto a las imputadas mediante Random Forest y k-vecinos más cercanos.

Se trabajó con un nivel de confianza del 95%, lo que corresponde a α = 0.05, valor utilizado en los análisis estadísticos. El p-value indica la probabilidad de obtener un resultado mayor o menor al observado; en base al mismo se decide rechazar o no la hipótesis nula (H0). En la investigación se aplicó la prueba de Kolmogorov-Smirnov, la cual indicó que no se rechaza la hipótesis nula (p menor que o igual a alfa), afirmando que no existe evidencia de diferencia significativa entre las distribuciones.

Existe suficiente evidencia para no rechazar la hipótesis nula, es decir, no difieren significativamente las bases de datos originales con respecto a los imputados. La prueba Kolmogorov-Smirnov (tabla 3) aplicada a las distribuciones de las bases de datos imputadas por Random Forest y k-vecinos más cercanos, indica que son estadísticamente iguales a las originales.La variable X_5 correspondiente a Radiación solar global promedio imputada mediante el método de k-Vecinos más cercanos, se rechaza la hipótesis nula, es decir, que la distribución es estadísticamente diferente a la original.

Tabla 3

Resultados y regla de decisión Kolmogorov-Smirnov.

	Variable	Comparación	p.value	Conclusión
$\mathbf{X_1}$	Temperatura ambiental promedio	orig vs k-NN	0.971	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.992	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_2}$	Humedad relativa promedio	orig vs k-NN	0.921	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.498	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_3}$	Presión atmosférica promedio	orig vs k-NN	0.856	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.637	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_4}$	Radiación solar difusa promedio	orig vs k-NN	0.943	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.975	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_5}$	Radiación solar global promedio	orig vs k-NN	0.001	Rechazó H_0 ($p < 0.05$)
		orig vs RF	0.182	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_6}$	Temperatura del suelo nivel 1	orig vs k-NN	0.999	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.875	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_7}$	Temperatura del suelo nivel 2	orig vs k-NN	0.993	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	1.000	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_8}$	Temperatura del suelo nivel 3	orig vs k-NN	0.999	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.999	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_9}$	Temperatura del suelo nivel 4 promedio	orig vs k-NN	1.000	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	1.000	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_{10}}$	Temperatura del suelo nivel 5	orig vs k-NN	1.000	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	1.000	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_{11}}$	Temperatura del suelo nivel 6	orig vs k-NN	0.994	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.956	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_{12}}$	Temperatura del suelo nivel 7	orig vs k-NN	0.959	No se rechazó H_0 ($p \geq 0.05$)
		orig vs RF	0.737	No se rechazó H_0 ($p \geq 0.05$)
$\mathbf{X_{13}}$	Precipitación acumulada	No se imputó, 87.85% de datos faltantes		
$\mathbf{X_{14}}$	Recorrido del viento	No se imputó, 11% de datos faltantes		
$\mathbf{X_{15}}$	Dirección del viento	No se imputó, 11% de datos faltantes		
$\mathbf{X_{16}}$	Dirección de la racha de viento	No se imputó, 11% de datos faltantes		
$\mathbf{X_{17}}$	Hora de la racha de viento	No se imputó, 11% de datos faltantes		

$\mathbf{X_{18}}$	Minuto de la racha de viento	No se imputó, 11% de datos faltantes		
$\mathbf{X_{19}}$	Velocidad promedio del viento	No se imputó, 11% de datos faltantes		
$\mathbf{X_{20}}$	Sensación térmica promedio	orig vs k-NN	0.999	No se rechazó H0. (p >= 0.05)
		orig vs RF	0.914	No se rechazó H0. (p >= 0.05)

En la tabla 4 se visualizan los resultados obtenidos RMSE y MAE de la base de datos del año 2019, los cuales evidencian que el algoritmo de Random Forest es más efectivo con respecto al método de k-vecinos más cercanos en la imputación de datos meteorológicos faltantes. Observando así que RF arrojó valores de RMSE y MAE más pequeños, con lo cual se indica que tiene mayor eficiencia al momento de imputar datos, para obtener estos resultados se promediaron las métricas obtenidas de las 11 estaciones meteorológicas que monitorea el GEAA en el año 2019.

Tabla 4

Resumen las métricas de evaluación obtenidas

Variable	RMSE_RF	MAE_RF	RMSE_k-NN	MAE_k-NN	Mejor_RMSE	Mejor_MAE
$\mathbf{X_1}$	0.724	0.509	1.2	0.845	RF	RF
$\mathbf{X_2}$	4.735	3.414	6.64	4.564	RF	RF
$\mathbf{X_3}$	7.519	2.945	10.587	4.656	RF	RF
$\mathbf{X_4}$	69.333	27.458	99.545	43.026	RF	RF
$\mathbf{X_5}$	49.807	21.948	79.367	38.156	RF	RF
$\mathbf{X_6}$	0.506	0.28	1.304	0.797	RF	RF
$\mathbf{X_7}$	0.377	0.22	1.135	0.73	RF	RF
$\mathbf{X_8}$	0.395	0.244	0.883	0.577	RF	RF
$\mathbf{X_9}$	0.374	0.234	0.759	0.503	RF	RF
$\mathbf{X_{10}}$	0.328	0.23	0.517	0.365	RF	RF
$\mathbf{X_{11}}$	0.246	0.172	0.345	0.229	RF	RF
$\mathbf{X_{12}}$	0.252	0.183	0.334	0.222	RF	RF
$\mathbf{X_{13} - X_{19}}$			No imputado			
$\mathbf{X_{20}}$	1.813	1.398	2.075	1.584	RF	RF

Evaluación del tiempo y memoria de imputación

$\boldsymbol{H_0}$: El tiempo y la memoria empleados siguen una distribución normal.

$\boldsymbol{H_1}$: El tiempo y la memoria empleados no siguen una distribución normal.

La prueba de Shapiro-Wilk (tabla 5) aplicada a los tiempos y memoria empleadas en la imputación muestra resultados diferenciados. Tanto el tiempo de usuario como el tiempo transcurrido en ambos métodos (RF y k-NN) siguen una distribución normal, al igual que el tiempo de sistema en RF y la

memoria Vcells en RF. Sin embargo, en el caso de tiempo de sistema en k-NN, memoria Vcells en k-NN y memoria Ncells en ambos métodos, los resultados evidencian que no se cumple la normalidad. Por lo tanto, en el siguiente análisis se requiere la aplicación de pruebas tanto paramétricas como no paramétricas para contrastar de manera adecuada los resultados.

Tabla 5

Resultados y regla de decisión Shapiro-Wilk

Variable		p.value	Conclusión
Tiempo_Usuario_Segs	RF	0.135	No se rechazó H_0 ($p \geq 0.05$)
	k-NN	0.502	No se rechazó H_0 ($p \geq 0.05$)
Tiempo_Sistema_Segs	RF	0.058	No se rechazó H_0 ($p \geq 0.05$)
	k-NN	0.003	Rechazó H_0 ($p \geq 0.05$)
Tiempo_Transcurrido_Segs	RF	0.077	No se rechazó H_0 ($p \geq 0.05$)
	k-NN	0.095	No se rechazó H_0 ($p \geq 0.05$)
Memoria_Vcells_MB	RF	0.146	No se rechazó H_0 ($p \geq 0.05$)
	k-NN	0.029	Rechazó H_0 ($p \geq 0.05$)
Memoria_Ncells_MB	RF	0.002	Rechazó H_0 ($p \geq 0.05$)
	k-NN	0.003	Rechazó H_0 ($p \geq 0.05$)

$\boldsymbol{H_0}$: No existe diferencia significativa sobre tiempo y memoria implementada en la imputación de datos al utilizar RF y k-NN.

$\boldsymbol{H_1}$: Existe diferencia significativa sobre tiempo y memoria implementada en la imputación de datos al utilizar RF y k-NN.

Existe suficiente evidencia para rechazar la hipótesis nula (H_0), es decir, la comparación de medias entre RFy k-NN (tabla 6) para el tiempo y la memoria en la imputación de datos muestra diferencias estadísticamente significativas en todas las variables. Los resultados sugieren que k-NN requiere un tiempo de ejecución menor, pero ligeramente mayor memoria, mientras que Random Forest requiere más tiempo con menor memoria. Por lo tanto, se debe considerar el costo computacional al elegir el método de imputación.

Tabla 6

Resultados y regla de decisión comparación de medias

Comparación		Método	p_value	Conclusión
Tiempo_Usuario_Segs	RF vs k-NN	t-Student para muestras independientes	0.0005	Rechazó H_0 ($p \geq 0.05$)
Tiempo_Sistema_Segs	RF vs k-NN	Wilcoxon	0.0010	Rechazó H_0 ($p \geq 0.05$)
Tiempo_Transcurrido_Segs	RF vs k-NN	t-Student para muestras independientes	0.0009	Rechazó H_0 ($p \geq 0.05$)

Memoria_Vcells_MB	RF vs k-NN	Wilcoxon	0.0010	Rechazó H_0 ($p \geq 0.05$)
Memoria_Ncells_MB	RF vs k-NN	Wilcoxon	0.0010	Rechazó H_0 ($p \geq 0.05$)

Los resultados de la tabla 7 indican que el método k-vecinos más cercanos demoró 33.3 segundos en promedio en imputar, siendo menor que 486.37 al momento de procesar el código de Random Forest. Se observó que k-NN utilizó mayor memoria virtual 8.41 MB que RF 7.11 MB, siendo mínima la diferencia. La memoria para objetos no vectoriales (Ncells), al ejecutar los dos métodos es de ~2.88 MB, siendo la misma.

Si la investigación se basara en la eficiencia computacional, evidentemente k-NN sería el mejor método de imputación, puesto que logra una imputación en menor tiempo con un consumo de memoria muy similar al de RF.

Tabla 7

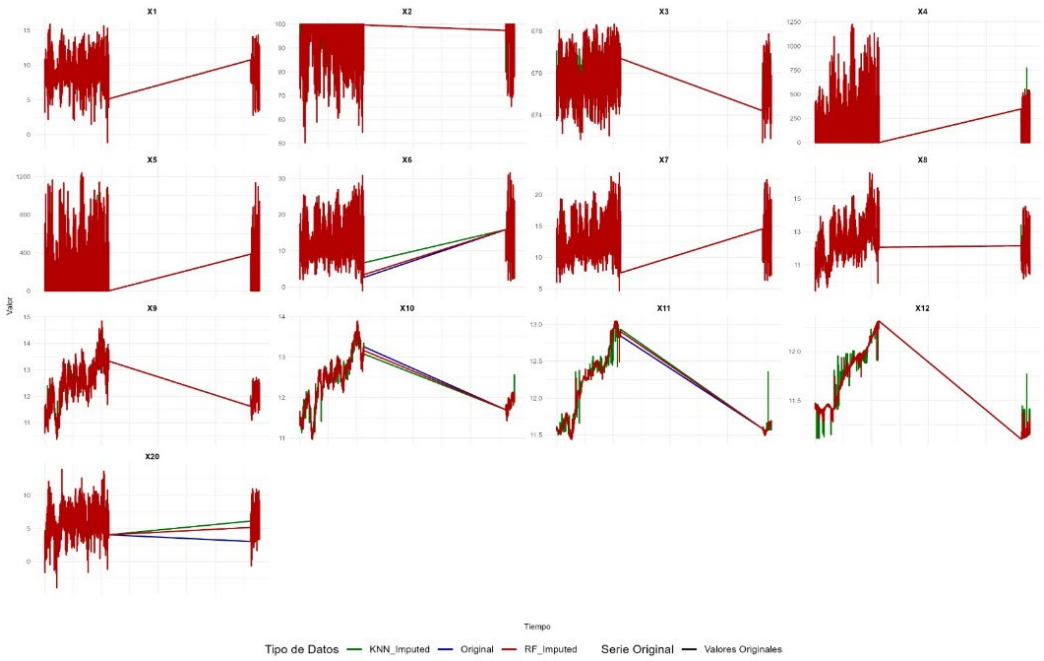
Tiempo de imputación / memoria virtual.

Metodo_Imputacion	Tiempo usuario segundos	Tiempo sistema segundos	Tiempo transcurrido segundos	Memoria Vcells MB	Memoria Ncells MB
Random Forest (RF)	286.047	1.025	486.375	7.107	2.882
K-Nearest Neighbors (KNN)	20.012	0.135	33.3	8.412	2.892

La figura 6 indica que los datos imputados tienen la misma tendencia en su distribución con respecto a los originales, ratificando visualmente lo obtenido en la prueba de Kolmogorov-Smirnov (tabla 3) que las distribuciones son estadísticamente iguales.

Figura 6

Análisis de datos imputados



4. Discusión

La ausencia de datos en registros meteorológicos representa una limitación frecuente que afecta directamente la calidad de los análisis climáticos y compromete la confiabilidad de los resultados obtenidos a partir de ellos. En este estudio se evaluaron dos métodos de imputación: Random Forest (RF) y k-Nearest Neighbors (k-NN), con el propósito de identificar cuál de ellos ofrece mejores condiciones para la reconstrucción de registros faltantes en las estaciones meteorológicas de la provincia de Chimborazo. Los resultados indicaron que RF superó a k-NN en términos de precisión, al obtener de manera consistente valores de RMSE y MAE más bajos en la mayoría de las variables analizadas. Este hallazgo se encuentra en línea con lo reportado por Breiman (2001) y Qi (2012), quienes destacan a RF como un algoritmo con alta capacidad para manejar relaciones no lineales y trabajar de manera eficiente con datos de gran dimensión.

El análisis comparativo también evidenció que, aunque k-NN demandó menos tiempo de ejecución y un uso ligeramente menor de memoria, su desempeño fue menos confiable en la imputación de los datos. Estos resultados reafirman las observaciones de Galván y Medina (2007), quienes describen a k-NN como un método computacionalmente eficiente, pero con limitaciones importantes cuando se enfrenta a grandes volúmenes de información o a estructuras de datos con patrones complejos. En este sentido, aunque k-NN puede resultar atractivo en entornos donde el costo computacional es una prioridad, su menor capacidad de ajuste frente a la variabilidad de los datos limita su utilidad en aplicaciones que requieren alta precisión, como es el caso de la predicción y análisis climático.

De manera complementaria, Velastegui y Horna (2023) concluyeron que, al aplicar distintos métodos de imputación como MICE, Hot Deck y RF en variables meteorológicas específicas, RF mostró un rendimiento superior en términos de exactitud y estabilidad de los resultados. Los hallazgos de este estudio refuerzan esa conclusión, pues RF no solo ofreció menores márgenes de error, sino que también mantuvo una estructura más coherente en la reconstrucción de las series temporales, lo que confirma su ventaja frente a métodos más simples.

Asimismo, los resultados permiten afirmar que el valor de RF no reside únicamente en su precisión, sino también en su robustez estadística. A diferencia de k-NN, que depende de la correcta elección de parámetros como el número de vecinos (k), RF integra múltiples árboles de decisión generados a partir de muestras bootstrap. Este mecanismo introduce diversidad en el modelo y reduce el sobreajuste, lo que le otorga una mayor capacidad para adaptarse a conjuntos de datos heterogéneos y con valores atípicos. Dichas características han sido documentadas en investigaciones de diversas áreas, como la predicción de enfermedades o la reconstrucción de series temporales complejas (Céspedes, 2022; Bashir et al., 2024). Esto evidencia que, a pesar de requerir un mayor consumo de recursos computacionales, RF ofrece un balance más favorable entre precisión y estabilidad, lo que lo convierte en una herramienta confiable para la imputación de datos meteorológicos incompletos.

Otro aspecto relevante es que, al analizar la aplicabilidad de los métodos, se observa que k-NN puede ser más útil en escenarios con limitaciones tecnológicas o en bases de datos pequeñas, donde su rapidez y bajo costo computacional compensan su menor exactitud. Sin embargo, cuando el objetivo es garantizar la calidad de los registros y minimizar los sesgos en los análisis, RF emerge como la opción más adecuada. Esto coincide con las conclusiones de Silva (2024), quien señaló que los métodos más robustos tienden a superar a los algoritmos simples en la imputación de series de precipitación, incluso en condiciones de alta complejidad.

En el contexto específico de Chimborazo, donde los registros meteorológicos presentan un porcentaje elevado de vacíos, la selección de un método de imputación robusto resulta determinante

para la generación de información confiable. Los resultados de este estudio permiten sostener que RF, pese a su mayor costo computacional, garantiza una mejor calidad en los registros reconstruidos y contribuye de manera más efectiva a la fiabilidad de los análisis climáticos. De este modo, la elección del método no debería guiarse únicamente por el tiempo de procesamiento o el uso de memoria, sino por un equilibrio entre eficiencia y precisión, siendo RF la alternativa más sólida para aplicaciones que requieren información de alta calidad.

Finalmente, al contrastar los hallazgos con lo reportado en la literatura, se observa una tendencia consistente: los métodos basados en enfoques más complejos, como RF, logran superar a técnicas tradicionales como k-NN en términos de exactitud, resistencia al ruido y capacidad de generalización. Esta convergencia entre la evidencia empírica y los estudios previos no solo valida los resultados obtenidos, sino que también abre la posibilidad de replicar la aplicación de RF en otras regiones que enfrentan problemas similares de incompletitud en sus registros meteorológicos. Así, se confirma que la imputación con RF no solo mejora la calidad de los datos, sino que también fortalece la toma de decisiones en escenarios donde la información climática precisa es esencial.

5. Conclusiones

El número de registros utilizados para este estudio fue en promedio de 8610 por cada una de las 13 variables, lo que permitió validar la metodología aplicada para la imputación de datos. Se comprobó que Random Forest, implementado en el software estadístico R, generó distribuciones equivalentes a las originales, lo que confirma su efectividad en la reconstrucción de registros meteorológicos incompletos.

Los análisis estadísticos (RMSE y MAE) mostraron que Random Forest ofrece mayor precisión que k-vecinos más cercanos, mientras que en términos de tiempo y memoria k-NN resultó más eficiente. Sin embargo, la capacidad de RF para reducir errores de predicción y manejar la complejidad de los datos lo convierte en el método más adecuado para garantizar la integridad de la información meteorológica.

Referencias

Amat, J. (2020). *Árboles de decisión, random forest, gradient boosting y C5.0.* RPubs by RStudio. https://rpubs.com/joaquin_ar/255596

Bashir, N., Mir, A., Daud, A., Rafique, M., & Bukhari, A. (2024). Time Series Reconstruction With Feature-Driven Imputation: A Comparison of Base Learning Algorithms. *IEEE Access*, *12*, 85511-85530. https://ieeexplore.ieee.org/abstract/document/10559977

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. https://doi.org/10.1023/A:1010933404324

Céspedes, B. (2022). *Aplicación del Algoritmo "Random Forest" para un modelo de clasificación sobre la tenencia de anemia de niños del Perú* [Tesis doctoral, Universidad Nacional del Santa]. Repositorio Institucional Digital. https://repositorio.uns.edu.pe/handle/20.500.14278/4007

Díaz-Uriarte, R., & Alvarez, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*(3). https://doi.org/10.1186/1471-2105-7-3

Dimitri, G., Cappelli, I., Scarselli, F., Fort, A., & Gori, M. (2024). Graph neural networks for missing data imputation in time series from meteorological sensors. In *Proceedings of IEEE MetroXRAINE 2024* (pp. 1242–1247). https://doi.org/10.1109/metroxraine62247.2024.10796616

Economou, T., Lazoglou, G., Tzyrkalli, A., Constantinidou, K., & Lelieveld, J. (2023). A data integration framework for spatial interpolation of temperature observations using climate model data. *PeerJ, 11,* e14519. https://doi.org/10.7717/peerj.14519

Galván, M., & Medina, F. (2007). *Imputación de datos: teoría y práctica* (Estudios Estadísticos 4755). Comisión Económica para América Latina y el Caribe (CEPAL). https://ideas.repec.org/p/ecr/col027/4755.html

Han, S., Kim, H., & Lee, Y. (2020). Double random forest. *Machine Learning, 109*, 1569–1586. https://doi.org/10.1007/s10994-020-05889-1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The Elements of Statistical Learning* (pp. 587-604). Springer. https://doi.org/10.1007/978-0-387-84858-7_15

Hayawi, K., Shahriar, S., & Hacid, H. (2024). Imputación de datos climáticos y mejora de la calidad mediante datos satelitales. *Revista de Ciencia de Datos y Sistemas Inteligentes*, *3* (2), 87-97. https://doi.org/10.47852/bonviewJDSIS42022857

Hernández-Sampieri, R., & Mendoza, C. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw Hill Education. https://doi.org/10.22201/fesc.20072236e.2019.10.18.6

Hou, T., Wu, L., Zhang, X., Wang, X., & Huang, J. (2023, November). STA-Net: Reconstruct Missing Temperature Data of Meteorological Stations Using a Spatiotemporal Attention Neural Network. In *International Conference on Neural Information Processing* (pp. 29-52). Singapore: Springer Nature Singapore. https://link.springer.com/chapter/10.1007/978-981-99-8126-7_3

Lacourly, N. (2012). Estadística multivariada. Ebooks Patagonia – J. C. Sáez Editor.

Qi, Y. (2012). *Random Forest for Bioinformatics*. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning* (pp. 307-323). Springer. https://doi.org/10.1007/978-1-4419-9326-7_11

Rizal, M., Wigena, A., & Afendi, F. (2022). Time series imputation using VAR-IM (case study: Weather data in meteorological station of Citeko). *BAREKENG: Journal of Mathematics and its applications*, *16*(4), 1373-1384. https://doi.org/10.30598/barekengvol16iss4pp1373-1384

Saied, M., & Guirguis, S. (2025). Explainable artificial intelligence for botnet detection in internet of things. *Scientific Reports, 15*, 7632. https://doi.org/10.1038/s41598-025-90420-6

Schonlau, M., & Zou, R. (2020). The random forest algorithm for statistical learning. *Stata Journal, 20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Silva, L. (2024). Comparación y Evaluación de Métodos para la Imputación de Precipitación Faltante. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(5), 11486-11501. https://doi.org/10.37811/cl_rcm.v8i5.14538

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Velastegui, E., & Horna, E. (2023). *Comparación de técnicas de relleno de datos faltantes de la variable velocidad de viento de los años 2014 al 2021* [Tesis de grado, Escuela Superior Politécnica de Chimborazo]. https://dspace.espoch.edu.ec/items/cb89c4e1-8554-447b-8f75-926704c44060

Yarupaita, B. (2021). *Modelación espacial de la susceptibilidad a incendios forestales en la región Junín utilizando el algoritmo Random Forest* [Tesis de maestría, Universidad Nacional del Centro del Perú]. Repositorio Institucional UNCP. http://hdl.handle.net/20.500.12894/7524

Zúñiga, I., & Crespo, E. (2021). *Meteorología y climatología.* UNED - Universidad Nacional de Educación a Distancia. https://dialnet.unirioja.es/servlet/libro?codigo=789631

Transparencia

Conflicto de interés

Los autores declaran que no existen conflictos de interés de naturaleza alguna como parte de la presente investigación.

Fuente de financiamiento

Los autores financiaron completamente la investigación.

Contribución de autoría

Laura Susana Naranjo Ordoñez: Conceptualización, metodología, software, validación, análisis formal, investigación, gestión de datos, visualización, redacción - preparación del borrador original, redacción - revisión y edición, financiamiento, administración del proyecto, recursos, supervisión.

Amalia Isabel Escudero Villa: Conceptualización, metodología, validación, análisis formal, investigación, gestión de datos, visualización, redacción - revisión y edición, financiamiento, recursos, supervisión.

Los autores contribuyeron activamente en el análisis de los resultados, revisión y aprobación del manuscrito final.