Ingeniería y Competitividad
Vol 28 (1) 2026 doi: 10.25100/iyc.v28i1.15506

Vol 28 (1)

ISSN
0123-3033
e- 2027-8284

Artículo Original

Predicción del acceso a la electricidad en hogares brasileños mediante aprendizaje automático

Prediction of electricity access in Brazilian households using machine learning

Leandro Scala da Rocha[1] João Bosco Gonçalves[1]

[1]Universidade Federal Rural do Rio de Janeiro (UFRRJ), Seropédica, Brazil.
[2]Universidade Federal do Espírito Santo (UFES), Vitória, Brazil.

Resumen

Introducción: A pesar de los avances hacia el acceso universal a la electricidad en Brasil, persisten focos de exclusión energética, especialmente en zonas rurales y en la región norte. Identificar con precisión estos territorios es fundamental para respaldar políticas públicas más efectivas y basadas en evidencia.
Objetivo: Proponer y evaluar un modelo de aprendizaje automático para estimar el porcentaje de hogares con acceso a la electricidad en Brasil, utilizando indicadores socioeconómicos del Índice de Desarrollo de Ciudades Sostenibles (IDCS).
Metodología: El estudio empleó un proceso de ciencia de datos que incluyó el preprocesamiento de los indicadores del IDCS, la selección de características y el ajuste de hiperparámetros. Se probaron diferentes algoritmos de aprendizaje supervisado, y su rendimiento se evaluó mediante métricas de error, en particular el RMSE y el MAPE. Tras una prueba comparativa, se seleccionó XGBoost como el modelo más adecuado.
Resultados: XGBoost obtuvo el mejor rendimiento predictivo, con un RMSE promedio de 3,42 y un MAPE de 1,78 %, lo que indica una alta precisión en la estimación del acceso a la electricidad. Las variables más relevantes fueron los ingresos de la población más pobre, la proporción de áreas forestales y naturales, y los indicadores relacionados con la educación juvenil.
Conclusión: Los resultados demuestran el potencial del aprendizaje automático como herramienta para apoyar el diagnóstico territorial y la formulación de políticas públicas orientadas a la universalización del acceso a la electricidad. El modelo propuesto ayuda a identificar los determinantes estructurales de la exclusión energética, proporcionando

Palabras clave: Electricidad; Accesibilidad a los servicios de salud; Mejora de los algoritmos de aprendizaje automático; Modelos teóricos; Estudios transversales.

Abstract

Introduction: Despite advances toward universal electricity access in Brazil, pockets of energy exclusion persist, particularly in rural areas and in the Northern region. Accurately identifying these territories is essential to support more effective, evidence-based public policies.
Objective: To propose and evaluate a machine learning model to estimate the percentage of households with access to electricity in Brazil, using socioeconomic indicators from the Sustainable Cities Development Index (SCDI).
Methodology: The study employed a data science pipeline including preprocessing of SCDI indicators, feature selection, and hyperparameter tuning. Different supervised learning algorithms were tested, with performance evaluated using error metrics, especially RMSE and MAPE. XGBoost was selected as the most suitable model after comparative testing.
Results: XGBoost achieved the best predictive performance, with an average RMSE of 3.42 and a MAPE of 1.78%, indicating high accuracy in estimating electricity access. The most relevant variables were income of the poorest population, the proportion of forested and natural areas, and indicators related to youth education.
Conclusion: The results demonstrate the potential of machine learning as a tool to support territorial diagnostics and the formulation of public policies aimed at universalizing electricity access. The proposed model helps identify structural determinants of energy exclusion, providing technical evidence to guide more targeted and efficient interventions.

Keywords: Electricity; Health Services Accessibility; Boosting Machine Learning Algorithms; Theoretical Models; Cross-Sectional Studies.

¿Cómo citar?

Scala da Rocha L, Bosco J. Predicción del acceso a la electricidad en hogares brasileños mediante aprendizaje automático. Ingeniería y Competitividad, 2026, 28(1)e-21315506

https://doi.org/10.25100/iyc.v28i1.15506

Recibido: 23/01/26
Revisado: 26/03/26
Aceptado: 15/04/26
Online: 22/04/26

Correspondencia

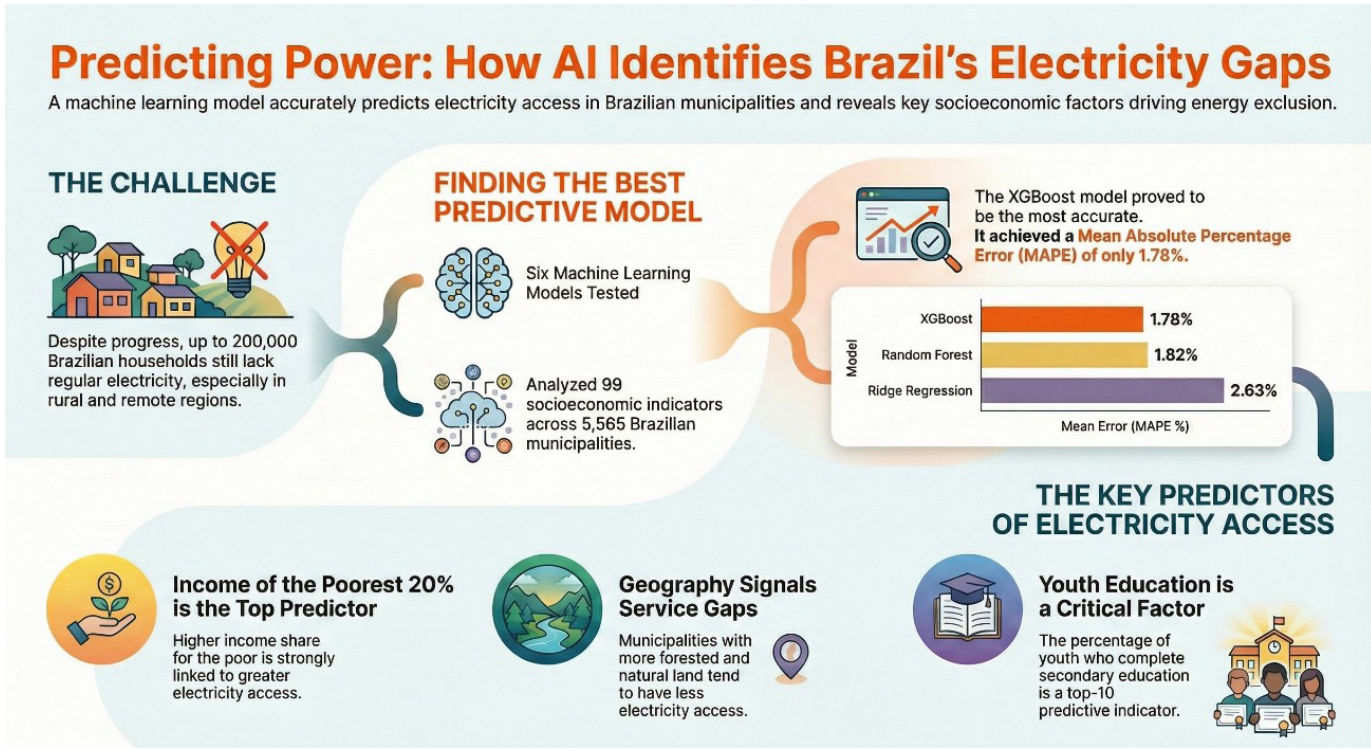
scala.leandro@ufrrj.br

¿Por qué se realizó?
El estudio se llevó a cabo para abordar una brecha clara y persistente: a pesar del progreso sustancial hacia el acceso universal a la electricidad en Brasil, persisten focos de exclusión energética, particularmente en zonas rurales y remotas. En este contexto, los autores buscaron llenar un vacío importante en la literatura, dada la falta de modelos predictivos robustos adaptados al contexto brasileño. Por consiguiente, el estudio buscó desarrollar un modelo capaz de estimar el acceso a la electricidad con base en indicadores socioeconómicos, lo que permite identificar territorios vulnerables y respalda el diseño de políticas públicas más efectivas y basadas en evidencia.

¿Cuáles fueron los resultados más relevantes?
Los hallazgos indican que el modelo XGBoost superó a todos los demás algoritmos probados, logrando una precisión predictiva superior, con un RMSE promedio de 3,42 y un MAPE de 1,78 %, además de una sólida capacidad explicativa. Más allá del desempeño del modelo, los resultados destacan que el acceso a la electricidad está estrechamente vinculado a factores socioeconómicos estructurales, particularmente los ingresos de la población más pobre, el nivel educativo de los jóvenes y la extensión de áreas forestales y naturales, lo que refleja los desafíos asociados con el aislamiento geográfico. El análisis SHAP reforzó aún más estos conocimientos al identificar no solo los indicadores más influyentes, sino también la dirección y magnitud de sus efectos en los distintos municipios.

¿Qué aportan?
El estudio realiza contribuciones relevantes en múltiples niveles. Científicamente, impulsa la literatura al integrar técnicas de aprendizaje automático con indicadores de desarrollo sostenible a escala nacional. Metodológicamente, ofrece un proceso analítico robusto, reproducible e interpretable que combina el rendimiento predictivo con la explicabilidad del modelo. Desde un punto de vista práctico, el modelo propuesto sirve como herramienta de inteligencia territorial, apoyando la toma de decisiones públicas al identificar áreas prioritarias y permitir una asignación más eficiente de recursos. En definitiva, el estudio transforma datos complejos en información útil, contribuyendo a la formulación de políticas dirigidas a lograr el acceso universal a la electricidad en Brasil.

Graphical Abstract



doi: 10.25100/iyc.v28i1.15506

Introducción

La universalización de la electrificación en Brasil representa uno de los pilares del desarrollo social, medioambiental y económico del país. La electricidad es un insumo fundamental para la promoción de derechos sociales básicos, como la salud, la educación y la seguridad alimentaria, además de ser un motor de productividad e inclusión social. En el contexto brasileño, donde aún existen focos de exclusión energética —especialmente en zonas remotas de la Amazonía Legal, así como en comunidades indígenas y quilombolas— garantizar el acceso universal a la electricidad es esencial para superar desigualdades históricas y regionales (1-3).

Estudios empíricos demuestran los efectos positivos de la electrificación sobre los ingresos, la educación, el empleo y la salud. Por ejemplo, Lipscomb, Mobarak y Barham (4) demostraron que, entre 1960 y 2000, la expansión de la red eléctrica en Brasil tuvo efectos significativos en el crecimiento de la renta per cápita, la reducción de la pobreza y la apreciación de los activos inmobiliarios. Estos hallazgos refuerzan que la electrificación va más allá de un derecho básico; es un instrumento de transformación estructural y multifactorial.

A pesar de los avances observados en las últimas décadas, todavía existen focos de exclusión energética en el país. Se estima que en 2021 hasta 200.000 hogares no tenían acceso regular a la electricidad (5).

En este contexto, es importante desarrollar un estimador de electrificación —es decir, un modelo predictivo capaz de estimar, basado en variables socioeconómicas y estructurales, la proporción de hogares con acceso a la electricidad. Un estimador robusto puede ayudar a mapear las áreas vulnerables, orientar políticas públicas e inversiones, y proporcionar una herramienta para el monitoreo continuo, con posibles aplicaciones en paneles de indicadores y plataformas de inteligencia territorial.

Aunque existen estudios centrados en analizar el acceso a la energía en contextos específicos—como los trabajos de Khandker, Barnes y Samad (6) en India; Santillán, Cedano y Martínez (7) en América Latina; y Wang et al. (8) en China—los modelos predictivos aplicados específicamente al contexto brasileño siguen siendo escasos. Los estudios nacionales, como el de Freitas y Oliveira (9), se centran predominantemente en análisis cualitativos o evaluaciones de impacto de programas como Luz para Todos, pero no abordan la construcción de estimadores basados en aprendizaje automático.

El Índice de Desarrollo de Ciudades Sostenibles (SCDI) fue creado por el Instituto Cidades Sustentáveis, basado en los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030, y proporciona datos organizados en 17 dimensiones, incluyendo educación, salud, infraestructuras, saneamiento, ingresos y acceso a la energía (10-11). Al abarcar múltiples ámbitos del desarrollo urbano y rural, los indicadores SCDI capturan indirectamente las condiciones estructurales que influyen en el acceso a la electricidad, permitiendo así una mayor robustez predictiva del modelo. Esto justifica la elección de indicadores SCDI como variables predictoras de la tasa de electrificación, dada su amplitud temática y relevancia territorial.

doi: 10.25100/iyc.v28i1.15506

La integración de modelos de aprendizaje automático con indicadores compuestos como el SCDI representa un enfoque novedoso para estimar, monitorizar y comprender la distribución de la electrificación en Brasil. Este enfoque contribuye a la formulación de políticas públicas más efectivas y al avance del conocimiento científico sobre los determinantes socioeconómicos de la transición energética en el país.

Este artículo pretende cubrir una carencia desarrollando un modelo predictivo para el porcentaje de hogares con acceso a la electricidad en municipios brasileños, utilizando técnicas de regresión supervisada. La característica distintiva de este trabajo es la aplicación integrada de una cadena de ciencia de datos, que abarca pasos desde la limpieza y normalización de datos hasta la selección y ajuste de hiperparámetros a través de múltiples modelos de regresión. El enfoque incluye el tratamiento de datos faltantes y variables colineales, la validación cruzada y la capacidad de explicar los resultados mediante valores SHAP (SHapley Additives ExPlanations), asegurando rigor estadístico y la interpretabilidad del modelo.

El artículo está organizado de la siguiente manera: la sección de Metodología presenta los datos utilizados, los pasos de preprocesamiento y la cadena de modelado. La sección de Resultados y Discusión informa sobre las métricas de rendimiento de los modelos probados y ofrece un análisis comparativo de los algoritmos. Finalmente, la Conclusión resume los principales hallazgos, analiza las limitaciones y propone direcciones para estudios futuros.

Metodología

El estudio utilizó los 100 indicadores componentes del SCDI para 5.565 municipios brasileños con valores disponibles para la variable objetivo, según lo proporcionado por el Instituto Cidades Sustentáveis (12) bajo el nombre SCDI-BR_2024. Uno de estos indicadores fue seleccionado como variable objetivo: el porcentaje de hogares con acceso a la electricidad (SDG7_2_ENRG). Los 99 indicadores restantes se utilizaron como variables predictoras.

Se analizaron seis modelos: Regresión de Cresta (13), Regresión de Lazo (14), Bosque Aleatorio (15), XGBoost (16), Regresión de Vector de Soporte (SVR) (17) y MLP (18-19).

El estudio se organizó en cuatro fases: (i) tratamiento de datos faltantes y variables colineales; (ii) criterios de selección de hiperparámetros para los modelos; (iii) selección del modelo con mejor rendimiento; y (iv) análisis de la importancia de las características y los valores SHAP para el modelo seleccionado.

Todos los análisis se implementaron en Python (versión 3.13.0), utilizando las librerías pandas y NumPy para manipulación de datos, scikit-learn para preprocesamiento, entrenamiento de modelos, ajuste de hiperparámetros y procedimientos de validación, XGBoost para modelado de aumento de gradiente y SHAP para análisis de explicabilidad de modelos. Se realizaron pruebas estadísticas usando SciPy, y se generaron visualizaciones de datos con Matplotlib y Seaborn.

El flujo de trabajo computacional incluía preprocesamiento, imputación de datos faltantes, diagnósticos de multicolinealidad, validación cruzada y comparación de modelos bajo una tubería reproducible.

doi: 10.25100/iyc.v28i1.15506

Tratamiento de datos faltantes y variables colineales

Los valores faltantes se imputaron utilizando el método de Imputación Multivariante por Ecuaciones Encadenadas (MICE). Este enfoque ajusta iterativamente modelos de regresión para cada indicador con valores faltantes, utilizando los indicadores restantes como predictores, hasta lograr la convergencia de los valores imputados. El método fue elegido por su capacidad para preservar relaciones multivariantes, evitando así sesgos introducidos por imputaciones simplistas como la sustitución por media o mediana (20-22).

Para evaluar la plausibilidad de las imputaciones, se aplicó la prueba de Kolmogorov–Smirnov (KS), comparando las distribuciones de los valores observados e imputados para cada indicador. Se eliminaron los indicadores cuya distancia KS superaba 0,2 con significación estadística ($p < 0,05$), ya que indicaban imputaciones inconsistentes (23-25).

La reducción de la multicolinealidad se llevó a cabo en dos etapas: (a) se eliminaron indicadores redundantes en pares con fuerte correlación lineal absoluta ($|r| > 0,9$) (26), manteniendo siempre el que tenía mayor correlación con la variable objetivo; (b) Se aplicó un análisis del Factor de Inflación de Varianza (VIF), eliminando iterativamente los indicadores hasta que todos los VIF presentados < 10 (27). La Figura 1 ilustra el proceso de reducción hasta 89.

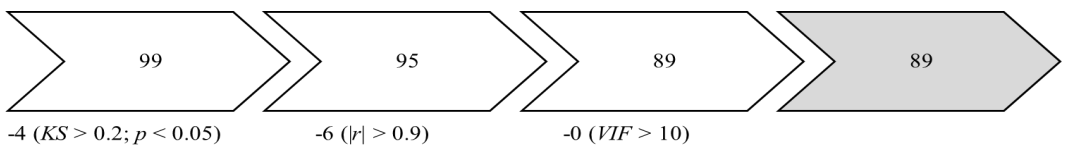


Figura 1. Proceso de reducción de indicadores. Fuente: Los autores

Criterios de selección de hiperparámetros para los modelos

Para evitar el sobreajuste y capturar patrones complejos en los indicadores, se seleccionaron hiperparámetros basándose en el valor medio más bajo del criterio de error cuadrático medio (MSE), utilizando la validación cruzada k-fold (k = 10) (28), con una división aleatoria de los datos en 80% para entrenamiento y 20% para pruebas. Los hiperparámetros seleccionados se muestran en la Tabla 1.

doi: 10.25100/iyc.v28i1.15506

Tabla 1. Hiperparámetros del modelo

Modelo	Hiperparámetros
Lazo	Alfa = 0.01
Cresta	Alfa = 1.000
Bosque Aleatorio	Número de árboles = 300, profundidad máxima de cada árbol = 20, número mínimo de muestras necesarias para dividir un nodo interno = 5
XGBoost	Número de árboles = 100, profundidad máxima de cada árbol = 6, tasa de aprendizaje = 0,1
SVR	C = 10, épsilon = 1, núcleo = lineal
MLP	Alfa = 10^{-4}, Tamaños de capas ocultas = (50, 50, 50, 50), Tasa de aprendizaje inicial = 0,1

Fuente: Los autores.

Selección del modelo con mejor rendimiento

Se compararon los valores medios de error cuadrático medio de raíz (RMSE) de 100 simulaciones independientes para cada modelo, tras verificar la normalidad de los resultados mediante la prueba de Shapiro-Wilk (29) y aplicar la prueba estadística de Kruskal–Wallis (30) a un nivel de significación del 5%. El modelo con la RMSE media más baja fue seleccionado como el modelo de mejor rendimiento.

Análisis de la importancia de las características y los valores SHAP para el modelo seleccionado

Los 10 indicadores más relevantes se identificaron en función de los coeficientes estandarizados del modelo con mejor rendimiento, permitiendo evaluar los determinantes más robustos de la variable objetivo. Además, se realizó un análisis de explicabilidad utilizando el método SHAP (31), aplicado al modelo seleccionado. Este método proporciona tanto la dirección como la magnitud de la contribución de cada indicador, permitiendo una explicación transparente de los resultados.

Resultados y discusión

El modelo XGBoost presentó los mejores valores medios de las métricas de rendimiento obtenidas de 100 simulaciones, como se muestra en la Tabla 2.

doi: 10.25100/iyc.v28i1.15506

Tabla 2. Valores medios de las métricas de rendimiento obtenidos de 100 simulaciones

Modelo	MSE	RMSE	MAE	R2	R2 Ajustado	MAPE (%)
Cresta	17.08	4.12	2.29	0.49	0.45	2.63
Lazo	16.83	4.09	2.35	0.50	0.45	2.68
Bosque Aleatorio	11.83	3.43	1.54	0.65	0.62	1.82
XGBoost	11.81	3.42	1.51	0.66	0.63	1.78
SVR	19.13	4.36	1.94	0.43	0.38	2.33
MLP	18.55	4.28	2.46	0.44	0.39	2.77

Fuente: Los autores.

Aunque el modelo de Bosque Aleatorio presentó métricas de rendimiento muy cercanas a las de XGBoost, este último fue seleccionado como modelo final porque logró los mejores resultados predictivos globales, como se muestra en la Tabla 2, especialmente en términos de RMSE media más baja (3,42), MAPE menor (1,78%) y R^2 ajustada más alta (0,63).

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje conjunto basado en árboles de decisión de gradiente boosting. El método construye árboles predictivos secuencialmente, donde cada nuevo árbol se entrena para minimizar los errores residuales producidos por los árboles anteriores. Sus principales ventajas incluyen mecanismos de regularización para reducir el sobreajuste, manejo eficiente de relaciones no lineales, robustez a multicolinealidad y alta escalabilidad computacional.

La Figura 2 muestra el valor medio de RMSE y el correspondiente intervalo de confianza del 95% tras 100 simulaciones para cada modelo. XGBoost presentó el valor medio RMSE más bajo (3,42).

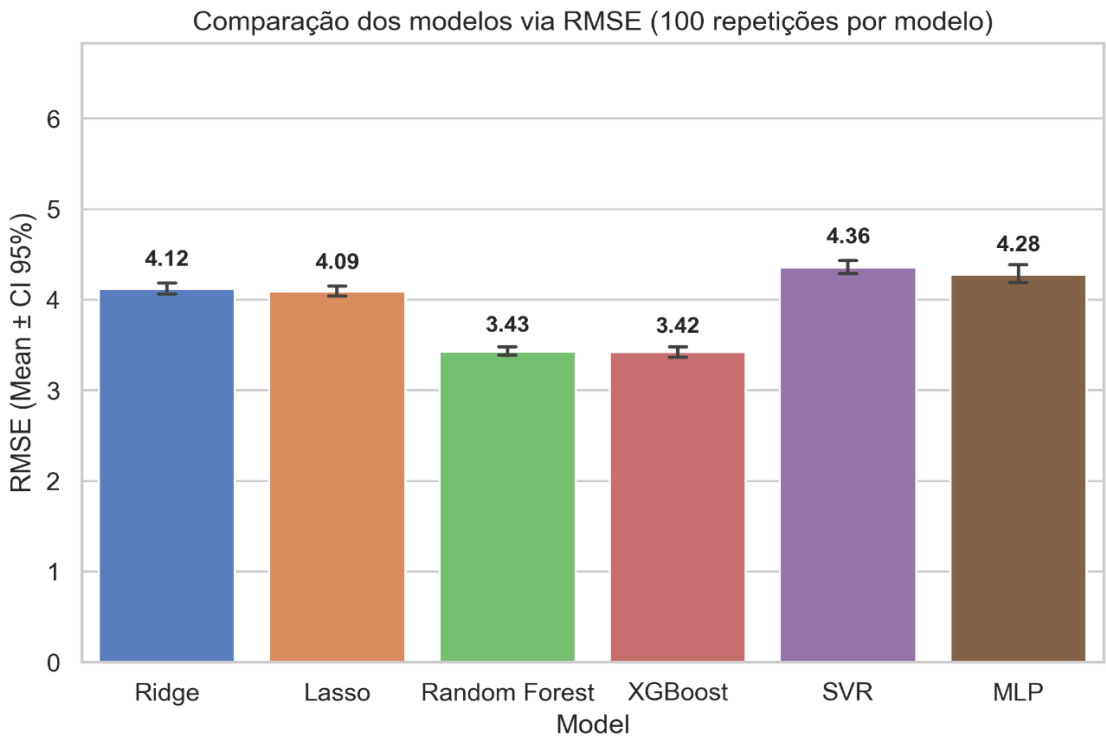


Figura 2. Valor medio e intervalo de confianza del 95% del RMSE para cada modelo, obtenidos a partir de 100 simulaciones. Fuente: Los autores

doi: 10.25100/iyc.v28i1.15506

La Tabla 3 presenta las estadísticas RMSE para cada modelo, obtenidas a partir de 100 simulaciones. XGBoost alcanzó la RMSE media más baja (3,42 ± 0,30), mientras que Random Forest mostró la desviación estándar más baja (0,25).

Tabla 3. Descripción estadística del RMSE para 100 simulaciones

Modelo	Promedio	Desviación estándar	Mínimo	25%	50%	75%	Máximo
Cresta	4.12	0.32	3.39	3.91	4.13	4.32	4.96
Lazo	4.09	0.29	3.39	3.92	4.13	4.29	4.69
Bosque Aleatorio	3.43	0.25	2.78	3.29	3.43	3.59	4.00
XGBoost	3.42	0.30	2.78	3.19	3.40	3.66	4.15
SVR	4.36	0.36	3.40	4.14	4.35	4.61	5.05
MLP	4.28	0.50	3.31	3.97	4.19	4.46	6.21

Fuente: Los autores.

La Tabla 3 muestra que Random Forest presentó una desviación estándar menor de RMSE (0,25) en comparación con XGBoost (0,30), lo que indica una estabilidad ligeramente mayor en las 100 simulaciones. Sin embargo, la diferencia en la desviación estándar fue pequeña y no comprometió la superioridad de XGBoost en términos de rendimiento medio.

La Figura 3 muestra la variabilidad del RMSE tras 100 simulaciones para cada modelo. Random Forest y XGBoost mostraron valores RMSE similares y fueron significativamente diferentes de los otros modelos, según la prueba de Kruskal–Wallis.

doi: 10.25100/iyc.v28i1.15506

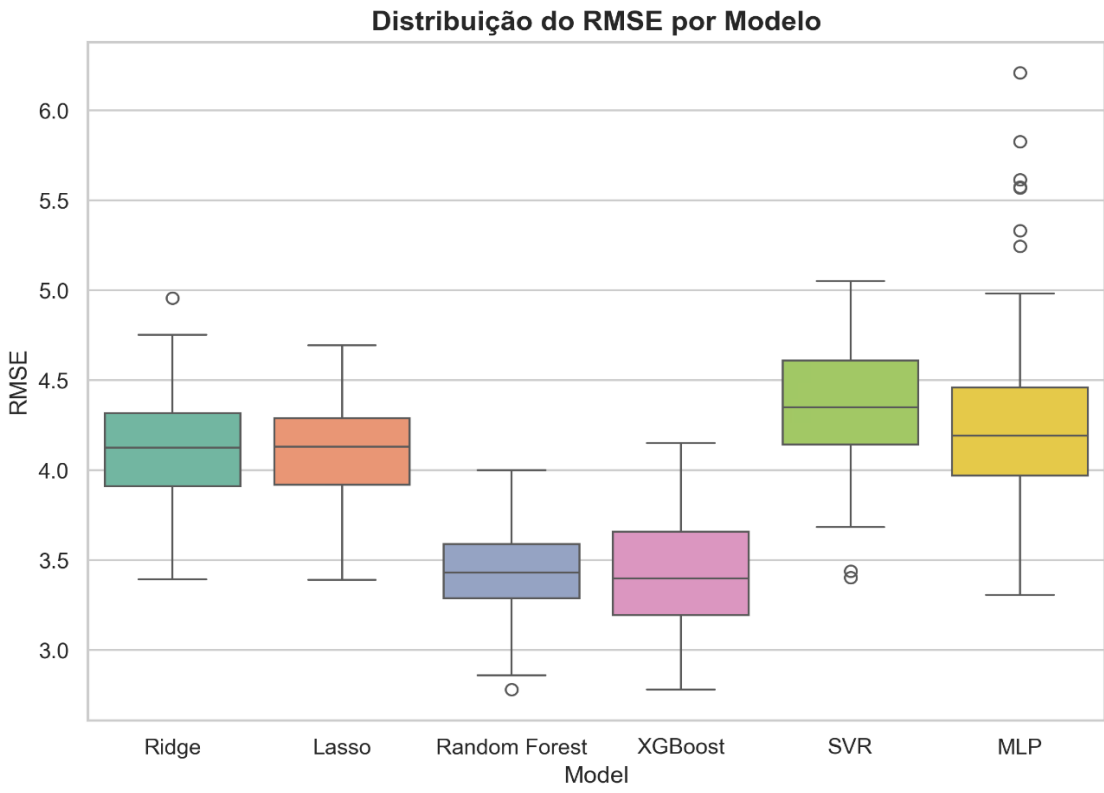


Figura 3. Variabilidad de RMSE obtenida a partir de 100 simulaciones. Fuente: Los autores.

Los 10 indicadores más importantes se muestran en la Tabla 4 y están principalmente relacionados con la educación y las desigualdades, según sus respectivos ODS (32).

doi: 10.25100/iyc.v28i1.15506

Tabla 4. Descripción de los indicadores más importantes

Código	ODS	Indicador
SDG10_1_PREN20	10 – Reducción de desigualdades	Ingresos municipales apropiados por el 20% más pobre (%)
SDG15_2_FLOR	15 – Vida en tierra	Hectáreas de áreas forestales y naturales per cápita
SDG4_17_JV_EM	4 – Educación de calidad	Jóvenes que completaron la educación secundaria superior antes de los 19 años (%)
SDG10_2_GINI	10 – Reducción de desigualdades	Coeficiente de Gini (IN)
SDG7_3_VNLENER	7 – Energía Asequible y Limpia	Vulnerabilidad energética
SDG6_6_SERV_AG	6 – Agua Limpia y Saneamiento	Población total atendida por el suministro de agua (%)
SDG4_3_TDI_EF_RP	4 – Educación de calidad	Tasa de distorsión por edad y por grado en la educación primaria – sistema público
SDG10_6_RND_NGR	10 – Reducción de desigualdades	Ratio de renta real media entre PP y BA
SDG16_3_M_ARM_FG	16 – Paz, Justicia e Instituciones Fuertes	Muertes relacionadas con armas de fuego (por cada 100.000 habitantes)
SDG4_27_R_M_D_EF	4 – Educación de calidad	Proporción entre el número de matrículas y profesores en educación primaria

Fuente: Los autores.

Los valores de las importancias relativas de los 10 indicadores principales se muestran en la Figura 4. Destacan los indicadores relacionados con los ingresos de la población más pobre (SDG10_1_ PREN20), áreas forestales y naturales (SDG15_2_FLOR) y educación juvenil (SDG4_17_JV_EM).

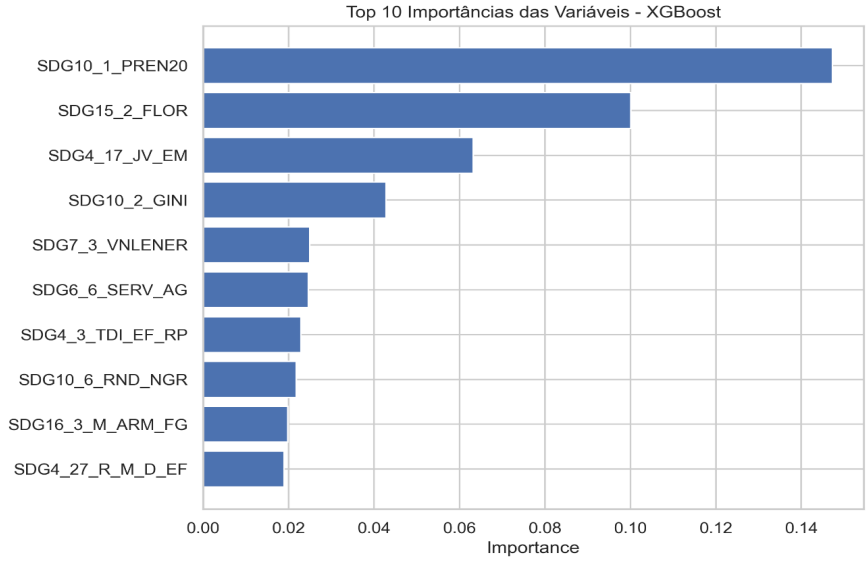


Figura 4. Las 10 principales importancias de los indicadores (XGBoost). Fuente: Los autores.

doi: 10.25100/iyc.v28i1.15506

La Tabla 5 muestra los 10 indicadores con los valores SHAP más altos, destacando aquellos relacionados con la educación y las desigualdades. Entre los principales indicadores está el relacionado con la vulnerabilidad energética (SDG7_3_VNLENER), que está directamente asociada con el acceso a la energía.

Tabla 5. Descripción de los indicadores con los valores SHAP más altos.

Código	ODS	Indicador
SDG15_2_FLOR	15 – Vida en tierra	Hectáreas de áreas forestales y naturales per cápita
SDG10_1_PREN20	10 – Reducción de desigualdades	Ingresos municipales apropiados por el 20% más pobre (%)
SDG4_17_JV_EM	4 – Educación de calidad	Jóvenes que completaron la educación secundaria superior antes de los 19 años (%)
SDG1_4_R_1_4_SM	1 – Sin pobreza	Población con ingresos de hasta 1/4 del salario mínimo (%)
SDG4_5_T_AN_15A	4 – Educación de calidad	Tasa de analfabetismo entre la población de 15 años o más (%)
SDG7_3_VNLENER	7 – Energía Asequible y Limpia	Vulnerabilidad energética
SDG4_3_TDI_EF_RP	4 – Educación de calidad	Tasa de distorsión por edad y por grado en la educación primaria – sistema público
SDG10_PVCM_NNN	10 – Reducción de desigualdades	Porcentaje de concejales PPI (negros, latinos e indígenas) en los Consejos Municipales (%)
SDG13_3_FCCLR	13 – Acción climática	Concentración de focos de incendios forestales
SDG8_2_OCP_INF	8 – Trabajo decente y crecimiento económico	Población empleada de 10 a 17 años (%)

Fuente: Los autores.

Los valores SHAP (Figura 5) complementaban los hallazgos indicando tanto la magnitud como la dirección del impacto de cada indicador. Se observa que áreas forestales y naturales per cápita (SDG15_2_FLOR) más grandes están negativamente asociadas con la variable objetivo (SDG7_2_ ENRG), mientras que los ingresos municipales apropiados por el 20% más pobre (SDG10_1_PREN20) están positivamente asociados con la variable objetivo. Esto demuestra que la privación energética se concentra en regiones remotas y entre poblaciones más pobres, corroborando los hallazgos de Galindo (1), Leduchowicz-Municio et al. (2), Pereira et al. (3), y Lipscomb, Mobarak y Barham (4).

doi: 10.25100/iyc.v28i1.15506

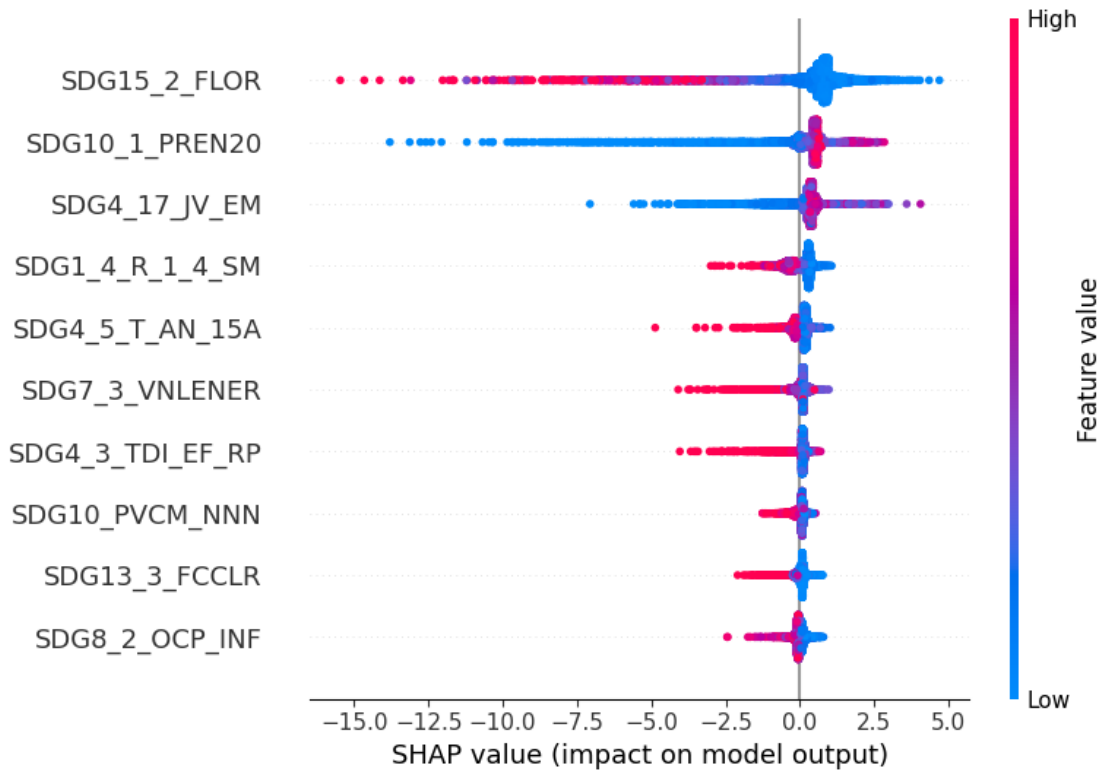


Figura 5. Top 10 de valores SHAP (XGBoost). Fuente: Los autores.

Aunque la Figura 5 indica una asociación negativa entre el indicador de áreas boscosas y naturales per cápita (SDG15_2_FLOR) y la variable objetivo, debe enfatizarse que los valores SHAP capturan relaciones estadísticas condicionales (dado el conjunto de covariables en el modelo) y no identifican causalidad. Así, el signo negativo significa que, manteniendo constantes los demás indicadores, valores más altos de áreas boscosas y naturales per cápita se asocian con valores más bajos de la variable objetivo; Esto no implica que la reducción de áreas boscosas y naturales conduzca a un aumento del porcentaje de hogares con acceso a la electricidad, ni viceversa.

Los indicadores presentados en el gráfico de importancia (Figura 4) y los mostrados en el gráfico SHAP (Figura 5) no coinciden completamente porque cada técnica aborda diferentes preguntas. El gráfico de importancia se construyó a partir de los valores medios de los coeficientes del modelo de regresión seleccionado, reflejando la contribución global de cada indicador según métricas lineales tradicionales. En cambio, los valores de SHAP consideran el impacto marginal de cada indicador en las predicciones del modelo para cada municipio, indicando tanto la magnitud como la dirección de los efectos.

Así, los conjuntos de indicadores pueden diferir porque las medidas de importancia priorizan la estabilidad, mientras que los valores SHAP reflejan exclusivamente la lógica interna del modelo de mejor rendimiento seleccionado (en este caso, XGBoost). Esta diferencia debe interpretarse de manera complementaria: la clasificación de importancia indica qué indicadores son consistentemente relevantes, mientras que SHAP muestra cómo estos indicadores influyen en la variable objetivo—positiva o negativamente—a nivel individual (municipal).

doi: 10.25100/iyc.v28i1.15506

Para mejorar la interpretabilidad territorial del modelo predictivo, la Figura 6 presenta la distribución geográfica del acceso a la electricidad previsto entre municipios brasileños. Esta representación espacial permite identificar grupos de vulnerabilidad, especialmente en regiones remotas y zonas con mayor privación socioeconómica.

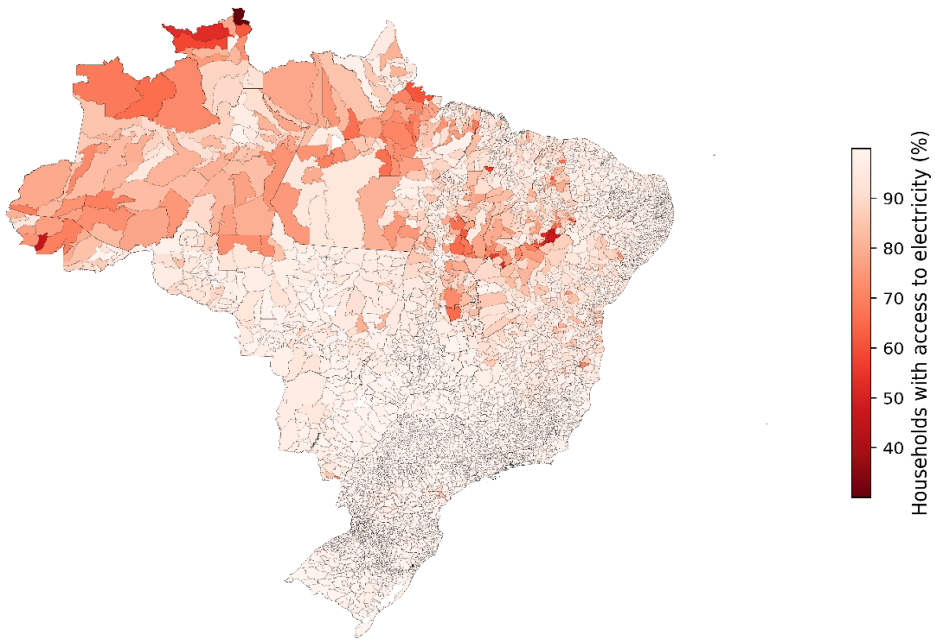


Figura 6. Distribución espacial del acceso eléctrico previsto entre municipios brasileños.
Fuente: Los autores.

Conclusión

Este estudio desarrolló un modelo predictivo para estimar el porcentaje de hogares con acceso a la electricidad en municipios brasileños, basado en variables socioeconómicas extraídas del Índice de Desarrollo Sostenible de Ciudades (SCDI). A través de una sólida cadena de datos de ciencia, el trabajo incluyó la imputación de datos multivariantes, la reducción de la multicolinealidad predictora, la selección de modelos de regresión, la afinación de hiperparámetros y el análisis de explicabilidad utilizando SHAP.

Entre los seis modelos evaluados—Ridge, Lasso, Random Forest, XGBoost, SVR y MLP—XGBoost logró el mejor rendimiento medio en términos de RMSE (3,42), R^2 ajustado (0,63) y MAPE (1,78%), demostrando efectividad en la estimación de la variable objetivo. El análisis estadístico confirmó diferencias significativas entre los modelos probados. Aunque Random Forest mostró un rendimiento similar, los mejores resultados predictivos promedio de XGBoost favorecieron su selección.

El análisis de la importancia predictiva reveló que los indicadores más relevantes eran los ingresos de la población más pobre, las áreas forestales y naturales, y la educación juvenil. Los 10 indicadores más importantes están relacionados con la desigualdad y la educación, mostrando

doi: 10.25100/iyc.v28i1.15506

coherencia con la literatura sobre exclusión energética y desigualdades estructurales. El análisis SHAP corrobora estos hallazgos, proporcionando explicabilidad tanto del modelo local como global, lo cual es especialmente útil para la formulación de políticas públicas específicas.

La principal contribución de este estudio es la propuesta de una herramienta replicable y escalable para el diagnóstico territorial del acceso a la electricidad, basada en aprendizaje automático y datos secundarios públicos. Mediante el uso de indicadores SCDI, el modelo permite predecir los niveles de electrificación municipal y comprender los factores estructurales asociados a este fenómeno. Este enfoque representa un avance respecto a estudios anteriores, más cualitativos o localizados.

Como limitación, el uso de datos agregados a nivel municipal puede ocultar desigualdades intraurbanas, además de la falta de validación con datos empíricos de campo. El estudio también se basó en datos transversales, lo que podría limitar la capacidad del modelo para la previsión a largo plazo.

Futuros estudios podrían integrar datos geoespaciales, series temporales o datos de sensores de consumo energético para aumentar la precisión y granularidad de las estimaciones. Los datos longitudinales pueden mejorar la capacidad predictiva del modelo a largo plazo.

Por tanto, se concluye que la modelización predictiva propuesta en este trabajo constituye una contribución relevante e innovadora al campo del análisis del acceso a la electricidad en Brasil, con posibles aplicaciones en la monitorización, priorización de políticas públicas y estudios sobre pobreza energética multidimensional.

Declaración de contribución de autoría de CrediT

Conceptualización - Ideas: Leandro Scala da Rocha. Conservación de datos: Derly Henao, Merly Daza, Michell Hernández. Análisis formal: Leandro Scala da Rocha. Investigación: Leandro Scala da Rocha. Metodología: Leandro Scala da Rocha. Gestión de Proyectos: João Bosco Gonçalves. Recursos: Leandro Scala da Rocha. Software: Leandro Scala da Rocha. Supervisión: João Bosco Gonçalves. Validación: Leandro Scala da Rocha e João Bosco Gonçalves. Guion - borrador original - Preparación: Aleandro Scala da Rocha e João Bosco Gonçalves. Escritura - revisión y edición - Preparación: Leandro Scala da Rocha e João Bosco GonçalvesFinanciación: no declarada.

Conflicto de intereses: no declarado. Consideraciones éticas: no declaradas.

Referencias

1. Galindo, MF. Acceso universal a la electricidad en zonas remotas: Construyendo un camino hacia la universalización en el Amazonas brasileño [Internet]. Estocolmo (SE): KTH Ingeniería Industrial y Gestión, Real Instituto de Tecnología; 2014 [citado 16 de diciembre de 2025]. Disponible en: https://www.diva-portal.org/smash/get/diva2:719200/fulltext01.pdf

2. Leduchowicz-Municio A, López-Gonzáles A, Domenech B, Ferrer-Martí L, Udaeta ME, Gimenes AL. Electrificación rural de última milla: Lecciones aprendidas de los programas de universalización en Brasil y Venezuela. Política energética. 2022; 167:113080.

https://doi.org/10.1016/j.enpol.2022.113080

doi: 10.25100/iyc.v28i1.15506

3. Pereira JD, Santos MA, Bandeira FD, Soares FI, Vieira TA. Electrificación en regiones remotas: Un análisis del programa More Light for Amazon. Energías. 2023; 16(12):4663.

https://doi.org/10.3390/en16124663

4. Lipscomb M, Mobarak AM, Barham T. Efectos del desarrollo de la electrificación: Evidencia de la colocación topográfica de centrales hidroeléctricas en Brasil. American Economic Journal: Economía aplicada. 2013; 5(2):200–231. Disponible en:

https://www.aeaweb.org/articles?id=10.1257/app.5.2.200

5. Brasil. Ministério de Minas e Energía (BR). Resenha energética brasileira: Ejercicio de 2022 [Internet]. Brasilia (DF): Ministério de Minas e Energía; 2023 [citado 16 de diciembre de 2025]. Disponible en: https://www.gov.br/mme/pt-br/assuntos/secretarias/sntep/publicacoes/resenha-energetica-brasileira/resenhas/resenha-energetica-2022.pdf/view

6. Khandker SR, Barnes DF, Samad HA. ¿Son también pobres en energía quienes tienen pocos ingresos? Pruebas de la India. Política energética. 2012; 47:1–12.

https://doi.org/10.1016/j.enpol.2012.02.028

7. Santillán OS, Cedano KG, Martínez M. Análisis de la pobreza energética en 7 países latinoamericanos utilizando un índice multidimensional de pobreza energética. Energías. 2020; 13(7):1608.

https://doi.org/10.3390/en13071608

8. Wang F, Geng H, Zha D, Zhang C. Pobreza energética multidimensional en China: Características de medición y disparidades espaciotemporales. Investigación de indicadores sociales. 2023; 168:45–78.

https://doi.org/10.1007/s11205-023-03129-2

9. Freitas GF, Oliveira ML. Un análisis del programa Luz para Todos del Gobierno Federal. Revista de Extensión y Estudios Rurales. 2017; 6(2):143–155.

https://doi.org/10.18540/rever622017143-155

10. Instituto de Ciudades Sostenibles. IDSC-BR: Introducción al Índice de Desarrollo Sostenible de Ciudades – Brasil [Internet]. São Paulo: Instituto de Ciudades Sostenibles; [fecha desconocida] [citado 15 de septiembre de 2025]. Disponible en:

https://idsc.cidadessustentaveis.org.br/introduction

11. Naciones Unidas, Brasil. Agenda 2030 para el Desarrollo Sostenible [Internet]. Brasilia (DF): Naciones Unidas Brasil; 15 de septiembre de 2015 [citado 16 de diciembre de 2025]. Disponible en:

https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustent%C3%A1vel

12. Instituto de Ciudades Sostenibles. IDSC-BR: Índice de Desarrollo Sostenible de Ciudades – Brasil [Internet]. São Paulo: Instituto de Ciudades Sostenibles; [fecha desconocida] [citado 15 de septiembre de 2025]. Disponible en:

https://www.cidadessustentaveis.org.br/paginas/idsc-br

13. Hoerl AE, Kennard RW. Regresión de crestas: Estimación sesgada para problemas no ortogonales. Tecnometría. 1970; 12(1):55–67.

https://doi.org/10.2307/1267351

14. Tibshirani R. Contracción y selección por regresión mediante el lazo. Journal of the Royal Statistical Society: Serie B. 1996; 58(1):267–288. Disponible en:

https://www.jstor.org/stable/2346178

15. Breiman L. Bosques aleatorios. Aprendizaje automático. 2001; 45(1):5–32.

https://doi.org/10.1023/A:1010933404324

16. Chen T, Guestrin C. XGBoost: Un sistema escalable de mejora de árboles. En las Actas de la 22ª Conferencia Internacional ACM SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos [Internet]; 13–17 de agosto de 2016; San Francisco, CA. Nueva York (NY): Asociación de Maquinaria de Computación; 2016 [citado 16 de diciembre de 2025]. p. 785–794).

https://doi.org/10.1145/2939672.2939785

17. Scikit-learn. sklearn.svm.SVR — documentación scikit-learn 1.3.0 [Internet]. [citado 15 de septiembre de 2025]. Disponible en:

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

18. Rosenblatt F. El perceptrón: Un modelo probabilístico para el almacenamiento y organización de la información en el cerebro. Revisión psicológica. 1958; 65(6):386–408. Disponible en:

https://www.ling.upenn.edu/courses/cogs501/Rosenblatt1958.pdf

19. Scikit-learn. sklearn.neural_network. MLPRegressor — documentación scikit-learn 1.3.0 [Internet]. [citado 15 de septiembre de 2025]. Disponible en:

https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

20. Scikit-learn. sklearn.impute. IterativeImputer — documentación scikit-learn 1.3.0 [Internet]. [citado 15 de septiembre de 2025]. Disponible en:

https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

21. Van Buuren S. Imputación flexible de datos faltantes. 2ª ed. Chapman y Hall/CRC; 2018.

22. Van Buuren S, Groothuis-Oudshoorn K. MICE: Imputación multivariante por ecuaciones encadenadas en R. Journal of Statistical Software. 2011; 45(3):1–67.

https://doi.org/10.18637/jss.v045.i03

23. Kolmogorov AN. Sobre la determinación empírica de una ley de distribución. Revista del Instituto Italiano de Actuarios. 1933; 4:83–91. Disponible en:

doi: 10.25100/iyc.v28i1.15506

http://digitale.bnc.roma.sbn.it/tecadigitale/giornale/CFI0353791/1933/unico/00000093

24. Massey FJ. La prueba Kolmogorov-Smirnov para comprobar la bonura del ajuste. Revista de la Asociación Estadística Americana. 1951; 46(253):68–78.

https://doi.org/10.1080/01621459.1951.10500769

25. Smirnov N. Tabla para estimar la bondad del ajuste de distribuciones empíricas. Anales de Estadística Matemática. 1948; 19(2):279–281. Disponible en:

https://www.jstor.org/stable/2236278

26. Akoglu H. Guía del usuario sobre coeficientes de correlación. Revista Turca de Medicina de Urgencias. 2018; 18(3):91–93.

https://doi.org/10.1016/j.tjem.2018.08.001

27. Marquardt DW. Inversos generalizados, regresión de crestas, estimación lineal sesgada y estimación no lineal. Tecnometría. 1970; 12(3):591–612.

https://doi.org/10.1080/00401706.1970.10488699

28. Burman P. Un estudio comparativo de la validación cruzada ordinaria, la validación cruzada en v-fold y los métodos repetidos de pruebas de aprendizaje. Biometrika. 1989; 76(3):503-514.

https://doi.org/10.2307/2336116

29. Shapiro SS, Wilk MB. Un análisis de la prueba de varianza para verificar la normalidad (muestras completas). Biometrika, v. 52, n. 3–4, p. 591–611, 1965.

https://doi.org/10.2307/2333709

30. Kruskal WH, Wallis W. A. Uso de rangos en el análisis de varianza de un solo criterio. Revista de la Asociación Estadística Americana. 1952; 47(260):583–621.

https://doi.org/10.2307/2280779

31. Lundberg SM., Lee S-I. Un enfoque unificado para interpretar las predicciones de modelos. En actas de la 31ª Conferencia Internacional sobre Sistemas de Procesamiento de Información Neuronal (NIPS 2017) [Internet]; 4-9 de diciembre de 2017; Long Beach, CA, EE. UU. Red Hook (NY): Curran Associates; 2017 [citado 16 de diciembre de 2025]. p. 4765–4774. Disponible en:

https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

32. Instituto de Ciudades Sostenibles. Agenda 2030 [Internet]. São Paulo: Instituto de Ciudades Sostenibles; [fecha desconocida] [citado 15 de septiembre de 2025]. Disponible en:

https://www.cidadessustentaveis.org.br/institucional/pagina/agenda2030

doi: 10.25100/iyc.v28i1.15506