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

Vol 28 (1)

ISSN
0123-3033
e- 2027-8284

Original Article

Prediction of electricity access in Brazilian households using machine learning

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

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.

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.

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 evidencia técnica para orientar intervenciones más específicas y eficientes.

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

How to cite?

Scala da Rocha L, Bosco J. Prediction of electricity access in Brazilian households using machine learning. Ingeniería y Competitividad, 2026, 28(1) e-21315506

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

Received: 23/01/26
Reviewed: 26/03/26
Accepted: 15/04/26
Online: 22/04/26

Correspondence

scala.leandro@ufrrj.br

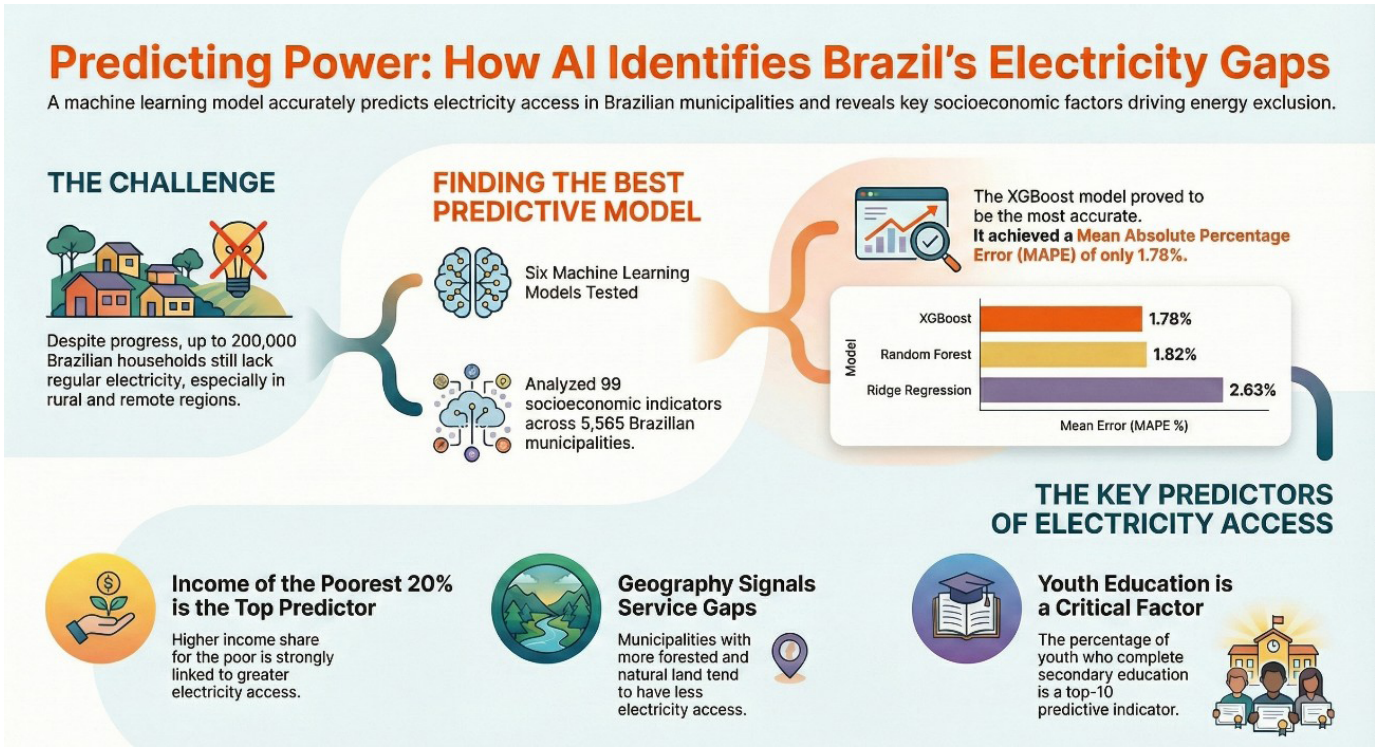
Spanish version

Why was it done?
The study was conducted to address a clear and persistent gap: despite substantial progress toward universal electricity access in Brazil, pockets of energy exclusion remain, particularly in rural and remote areas. In this context, the authors aimed to fill an important gap in the literature, given the lack of robust predictive models tailored to the Brazilian setting. Accordingly, the study sought to develop a model capable of estimating electricity access based on socioeconomic indicators, enabling the identification of vulnerable territories and supporting the design of more effective, evidence-based public policies.

What were the most relevant results?
The findings indicate that the XGBoost model outperformed all other tested algorithms, achieving superior predictive accuracy, with an average RMSE of 3.42 and a MAPE of 1.78%, alongside strong explanatory capacity. Beyond model performance, the results highlight that electricity access is closely linked to structural socioeconomic factors, particularly the income of the poorest population, youth educational attainment, and the extent of forested and natural areas—reflecting challenges associated with geographic isolation. The SHAP analysis further enhanced these insights by identifying not only the most influential indicators but also the direction and magnitude of their effects across municipalities.

What do they contribute?
The study makes relevant contributions at multiple levels. Scientifically, it advances the literature by integrating machine learning techniques with sustainable development indicators at a national scale. Methodologically, it offers a robust, reproducible, and interpretable analytical pipeline that combines predictive performance with model explainability. From a practical standpoint, the proposed model serves as a territorial intelligence tool, supporting public decision-making by identifying priority areas and enabling more efficient allocation of resources. Ultimately, the study translates complex data into actionable insights, contributing to the formulation of policies aimed at achieving universal electricity access in Brazil.

Graphical Abstract



doi: 10.25100/iyc.v28i1.15506

Introduction

The universalization of electrification in Brazil represents one of the pillars of the country's social, environmental, and economic development. Electricity is a fundamental input for the promotion of basic social rights, such as health, education, and food security, in addition to being a driver of productivity and social inclusion. In the Brazilian context, where pockets of energy exclusion still exist—especially in remote areas of the Legal Amazon, as well as in Indigenous and quilombola communities—ensuring universal access to electricity is essential to overcoming historical and regional inequalities (1-3).

Empirical studies demonstrate the positive effects of electrification on income, education, employment, and health. For example, Lipscomb, Mobarak, and Barham (4) showed that, between 1960 and 2000, the expansion of the electricity grid in Brazil had significant effects on per capita income growth, poverty reduction, and the appreciation of housing assets. These findings reinforce that electrification goes beyond a basic right; it is an instrument of structural and multifactorial transformation. Despite the advances observed in recent decades, pockets of energy exclusion still exist in the country. It is estimated that in 2021 up to 200,000 households did not have regular access to electricity (5).

In this context, the development of an electrification estimator—that is, a predictive model capable of estimating, based on socioeconomic and structural variables, the proportion of households with access to electricity—is important. A robust estimator can help map vulnerable areas, guide public policies and investments, and provide a tool for continuous monitoring, with potential applications in indicator dashboards and territorial intelligence platforms.

Although there are studies focused on analyzing access to energy in specific contexts—such as the works of Khandker, Barnes, and Samad (6) in India; Santillán, Cedano, and Martínez (7) in Latin America; and Wang et al. (8) in China—predictive models applied specifically to the Brazilian context remain scarce. National studies, such as that by Freitas and Oliveira (9), predominantly focus on qualitative analyses or impact evaluations of programs such as *Luz para Todos*, but do not address the construction of estimators based on machine learning.

The Sustainable Cities Development Index (SCDI) was created by the *Instituto Cidades Sustentáveis*, based on the Sustainable Development Goals (SDGs) of the 2030 Agenda, and provides data organized into 17 dimensions, including education, health, infrastructure, sanitation, income, and access to energy (10-11). By encompassing multiple spheres of urban and rural development, SCDI indicators indirectly capture structural conditions that influence access to electricity, thereby allowing greater predictive robustness of the model. This justifies the choice of SCDI indicators as predictor variables of the electrification rate, given their thematic breadth and territorial relevance.

The integration of machine learning models with composite indicators such as the SCDI represents a novel approach to estimating, monitoring, and understanding the distribution of electrification in Brazil. This approach contributes to the formulation of more effective public policies and to the advancement of scientific knowledge on the socioeconomic determinants of the energy transition in the country.

doi: 10.25100/iyc.v28i1.15506

This article aims to fill a gap by developing a predictive model for the percentage of households with access to electricity in Brazilian municipalities, using supervised regression techniques. The distinguishing feature of this work is the integrated application of a data science pipeline, encompassing steps from data cleaning and normalization to the selection and tuning of hyperparameters across multiple regression models. The approach includes the treatment of missing data and collinear variables, cross-validation, and the ability to explain results through SHAP (SHapley Additive exPlanations) values, ensuring statistical rigor and model interpretability.

The paper is organized as follows: the Methodology section presents the data used, the preprocessing steps, and the modeling pipeline. The Results and Discussion section reports on the performance metrics of the tested models and provides a comparative analysis of the algorithms. Finally, the Conclusion summarizes the main findings, discusses the limitations, and proposes directions for future studies.

Methodology

The study used the 100 component indicators of the SCDI for 5,565 Brazilian municipalities with available values for the target variable, as provided by the *Instituto Cidades Sustentáveis* (12) under the name SCDI-BR_2024. One of these indicators was selected as the target variable: the percentage of households with access to electricity (SDG7_2_ENRG). The remaining 99 indicators were used as predictor variables.

Six models were analyzed: Ridge Regression (13), Lasso Regression (14), Random Forest (15), XGBoost (16), Support Vector Regression (SVR) (17), and MLP (18-19).

The study was organized into four stages: (i) treatment of missing data and collinear variables; (ii) hyperparameter selection criteria for the models; (iii) selection of the best-performing model; and (iv) analysis of feature importance and SHAP values for the selected model.

All analyses were implemented in Python (version 3.13.0), using the libraries pandas and NumPy for data manipulation, scikit-learn for preprocessing, model training, hyperparameter tuning, and validation procedures, XGBoost for gradient boosting modeling, and SHAP for model explainability analysis. Statistical tests were conducted using SciPy, and data visualizations were generated with Matplotlib and Seaborn.

The computational workflow included preprocessing, missing data imputation, multicollinearity diagnostics, cross-validation, and model comparison under a reproducible pipeline.

Treatment of missing data and collinear variables

Missing values were imputed using the Multivariate Imputation by Chained Equations (MICE) method. This approach iteratively fits regression models for each indicator with missing values, using the remaining indicators as predictors, until convergence of the imputed values is achieved. The method was chosen for its ability to preserve multivariate relationships, thereby avoiding biases introduced by simplistic imputations such as mean or median substitution (20-22).

doi: 10.25100/iyc.v28i1.15506

To assess the plausibility of the imputations, the Kolmogorov–Smirnov (KS) test was applied, comparing the distributions of observed and imputed values for each indicator. Indicators whose KS distance exceeded 0.2 with statistical significance ($p < 0.05$) were removed, as they indicated inconsistent imputations (23-25).

The reduction of multicollinearity was carried out in two stages: (a) redundant indicators in pairs with strong absolute linear correlation ($|r| > 0.9$) were removed (26), always retaining the one with the highest correlation with the target variable; (b) Variance Inflation Factor (VIF) analysis was applied, iteratively removing indicators until all presented VIF < 10 (27). Figure 1 illustrates the reduction process down to 89.

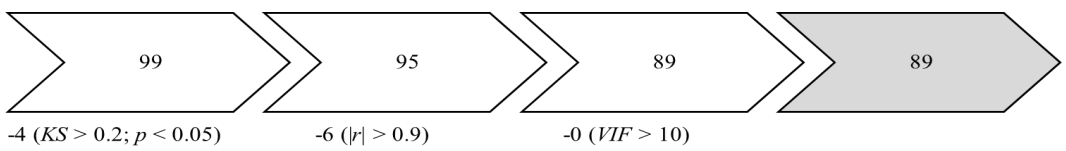


Figure 1. Indicator reduction process. Source: The authors

Hyperparameter selection criteria for the models

To avoid overfitting while still capturing complex patterns in the indicators, hyperparameters were selected based on the lowest mean value of the mean squared error (MSE) criterion, using k-fold cross-validation (k = 10) (28), with a random split of the data into 80% for training and 20% for testing. The selected hyperparameters are shown in Table 1.

Table 1. Model hyperparameters

Model	Hyperparameters
Lasso	Alpha = 0.01
Ridge	Alpha = 1,000
Random Forest	Number of trees = 300, Maximum depth of each tree = 20, Minimum number of samples required to split an internal node = 5
XGBoost	Number of trees = 100, Maximum depth of each tree = 6, Learning rate = 0.1
SVR	C = 10, epsilon = 1, kernel = linear
MLP	Alpha = 10^{-4}, Hidden layer sizes = (50, 50, 50, 50), Initial learning rate = 0.1

Source: The authors.

Selection of the best-performing model

The mean root mean square error (RMSE) values from 100 independent simulations were compared for each model, following verification of result normality using the Shapiro-Wilk test (29) and

doi: 10.25100/iyc.v28i1.15506

application of the Kruskal–Wallis statistical test (30) at a 5% significance level. The model with the lowest mean RMSE was selected as the best-performing model.

Analysis of feature importance and SHAP values for the selected model

The 10 most relevant indicators were identified based on the standardized coefficients of the best-performing model, allowing the assessment of the most robust determinants of the target variable. In addition, an explainability analysis was conducted using the SHAP method (31), applied to the selected model. This method provides both the direction and the magnitude of each indicator's contribution, enabling a transparent explanation of the results.

Results and Discussion

The XGBoost model presented the best average values of the performance metrics obtained from 100 simulations, as shown in Table 2.

Table 2. Mean values of the performance metrics obtained from 100 simulations

Model	MSE	RMSE	MAE	R^2	R^2 Adjusted	MAPE (%)
Ridge	17.08	4.12	2.29	0.49	0.45	2.63
Lasso	16.83	4.09	2.35	0.50	0.45	2.68
Random Forest	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

Source: The authors.

Although the Random Forest model presented performance metrics very close to those of XGBoost, the latter was selected as the final model because it achieved the best overall predictive results, as shown in Table 2, particularly in terms of lower mean RMSE (3.42), lower MAPE (1.78%), and higher adjusted R^2 (0.63).

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on gradient boosting decision trees. The method builds predictive trees sequentially, where each new tree is trained to minimize the residual errors produced by the previous trees. Its main advantages include regularization mechanisms to reduce overfitting, efficient handling of nonlinear relationships, robustness to multicollinearity, and high computational scalability.

Figure 2 shows the mean RMSE value and the corresponding 95% confidence interval after 100 simulations for each model. XGBoost presented the lowest mean RMSE value (3.42).

doi: 10.25100/iyc.v28i1.15506

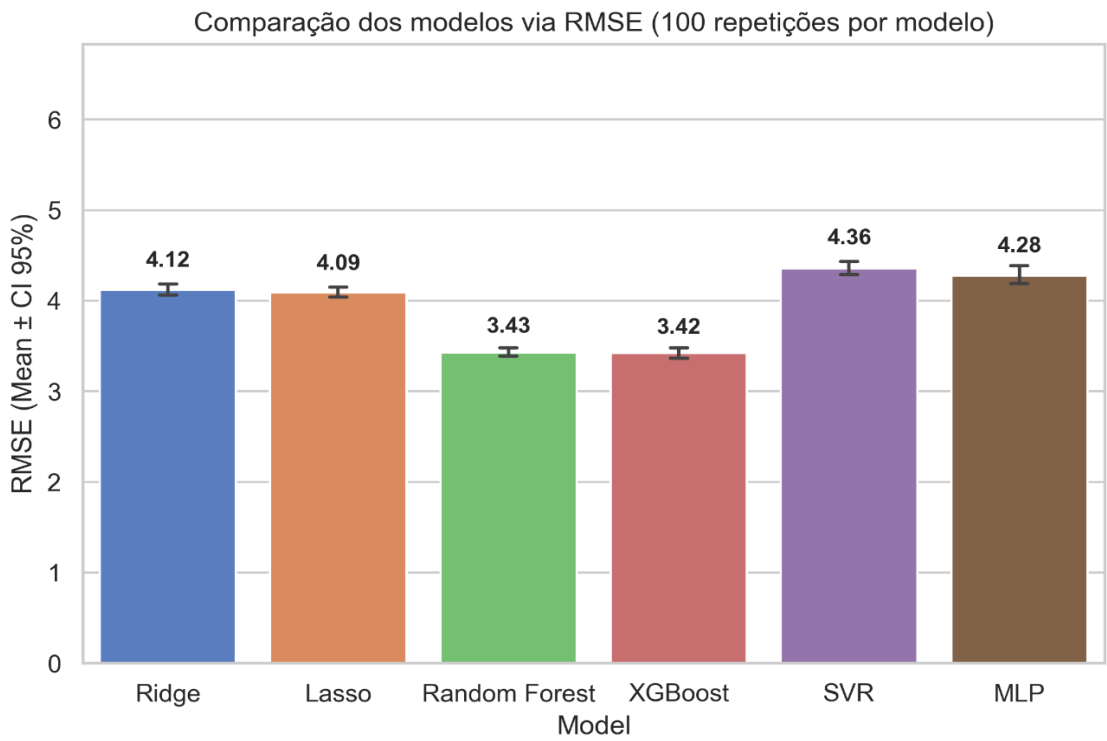


Figure 2. Mean value and 95% confidence interval of the RMSE for each model, obtained from 100 simulations. Source: The authors

Table 3 presents the RMSE statistics for each model, obtained from 100 simulations. XGBoost achieved the lowest mean RMSE (3.42 ± 0.30), while Random Forest showed the lowest standard deviation (0.25).

Table 3. Statistical description of RMSE for 100 simulations

Model	Mean	Standard deviation	Minimum	25%	50%	75%	Maximum
Ridge	4.12	0.32	3.39	3.91	4.13	4.32	4.96
Lasso	4.09	0.29	3.39	3.92	4.13	4.29	4.69
Random Forest	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

Source: The authors.

Table 3 shows that Random Forest presented a lower standard deviation of RMSE (0.25) compared with XGBoost (0.30), indicating slightly greater stability across the 100 simulations. However, the difference in standard deviation was small and did not compromise the superiority of XGBoost in terms of average performance.

doi: 10.25100/iyc.v28i1.15506

Figure 3 shows the variability of the RMSE after 100 simulations for each model. Random Forest and XGBoost exhibited similar RMSE values and were significantly different from the other models, according to the Kruskal–Wallis test.

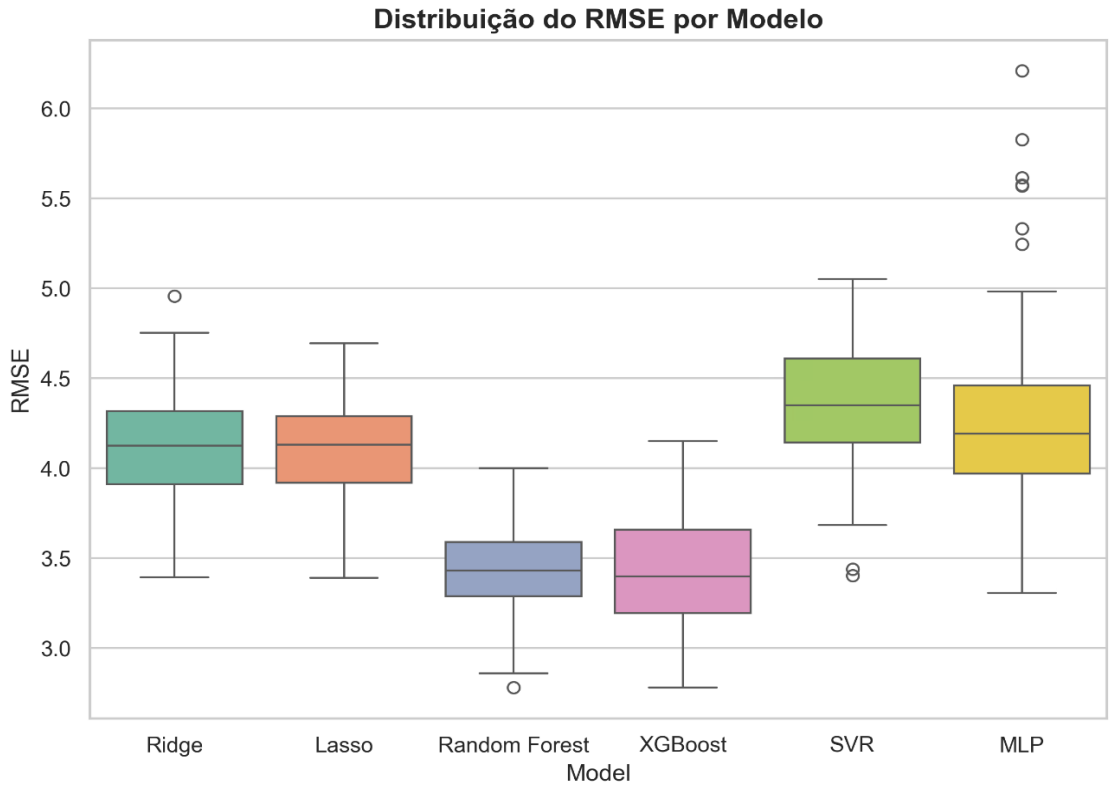


Figure 3. Variability of RMSE obtained from 100 simulations. Source: The authors.

The 10 most important indicators are shown in Table 4 and are mainly related to education and inequalities, as indicated by their respective SDGs (32).

doi: 10.25100/iyc.v28i1.15506

Table 4. Description of the most important indicators

Code	SDG	Indicator
SDG10_1_PREN20	10 – Reduced Inequalities	Municipal income appropriated by the poorest 20% (%)
SDG15_2_FLOR	15 – Life on Land	Hectares of forested and natural areas per capita
SDG4_17_JV_EM	4 – Quality Education	Youth who completed upper secondary education by age 19 (%)
SDG10_2_GINI	10 – Reduced Inequalities	Gini coefficient (IN)
SDG7_3_VNLENER	7 – Affordable and Clean Energy	Energy vulnerability
SDG6_6_SERV_AG	6 – Clean Water and Sanitation	Total population served by water supply (%)
SDG4_3_TDI_EF_RP	4 – Quality Education	Age–grade distortion rate in primary education – public system
SDG10_6_RND_NGR	10 – Reduced Inequalities	Ratio of average real income between PP and BA
SDG16_3_M_ARM_FG	16 – Peace, Justice, and Strong Institutions	Firearm-related deaths (per 100,000 inhabitants)
SDG4_27_R_M_D_EF	4 – Quality Education	Ratio between the number of enrollments and teachers in primary education

Source: The authors.

The values of the relative importances for the 10 main indicators are shown in Figure 4. Indicators related to the income of the poorest population (SDG10_1_PREN20), forested and natural areas (SDG15_2_FLOR), and youth education (SDG4_17_JV_EM) stand out.

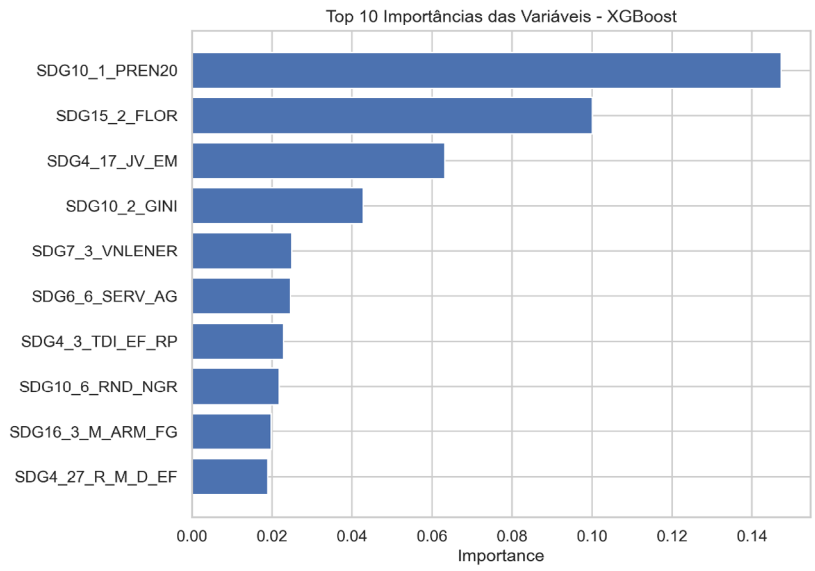


Figure 4. Top 10 Indicator Importances (XGBoost). Source: The authors.

doi: 10.25100/iyc.v28i1.15506

Table 5 shows the 10 indicators with the highest SHAP values, highlighting those related to education and inequalities. Among the main indicators is the one related to energy vulnerability (SDG7_3_VNLENER), which is directly associated with access to energy.

Table 5. Description of the indicators with the highest SHAP values.

Code	SDG	Indicator
SDG15_2_FLOR	15 – Life on Land	Hectares of forested and natural areas per capita
SDG10_1_PREN20	10 – Reduced Inequalities	Municipal income appropriated by the poorest 20% (%)
SDG4_17_JV_EM	4 – Quality Education	Youth who completed upper secondary education by age 19 (%)
SDG1_4_R_1_4_SM	1 – No Poverty	Population with income up to 1/4 of the minimum wage (%)
SDG4_5_T_AN_15A	4 – Quality Education	Illiteracy rate among the population aged 15 years or older (%)
SDG7_3_VNLENER	7 – Affordable and Clean Energy	Energy vulnerability
SDG4_3_TDI_EF_RP	4 – Quality Education	Age–grade distortion rate in primary education – public system
SDG10_PVCM_NNN	10 – Reduced Inequalities	Percentage of PPI (Black, Brown, and Indigenous) city councilors in Municipal Councils (%)
SDG13_3_FCCLR	13 – Climate Action	Concentration of wildfire hotspots
SDG8_2_OCP_INF	8 – Decent Work and Economic Growth	Employed population aged 10 to 17 years (%)

Source: The authors.

The SHAP values (Figure 5) complemented the findings by indicating both the magnitude and the direction of the impact of each indicator. It is observed that larger forested and natural areas per capita (SDG15_2_FLOR) are negatively associated with the target variable (SDG7_2_ENRG), while municipal income appropriated by the poorest 20% (SDG10_1_PREN20) is positively associated with the target variable. This demonstrates that energy deprivation is concentrated in remote regions and among poorer populations, corroborating the findings of Galindo (1), Leduchowicz-Municio et al. (2), Pereira et al. (3), and Lipscomb, Mobarak, and Barham (4).

doi: 10.25100/iyc.v28i1.15506

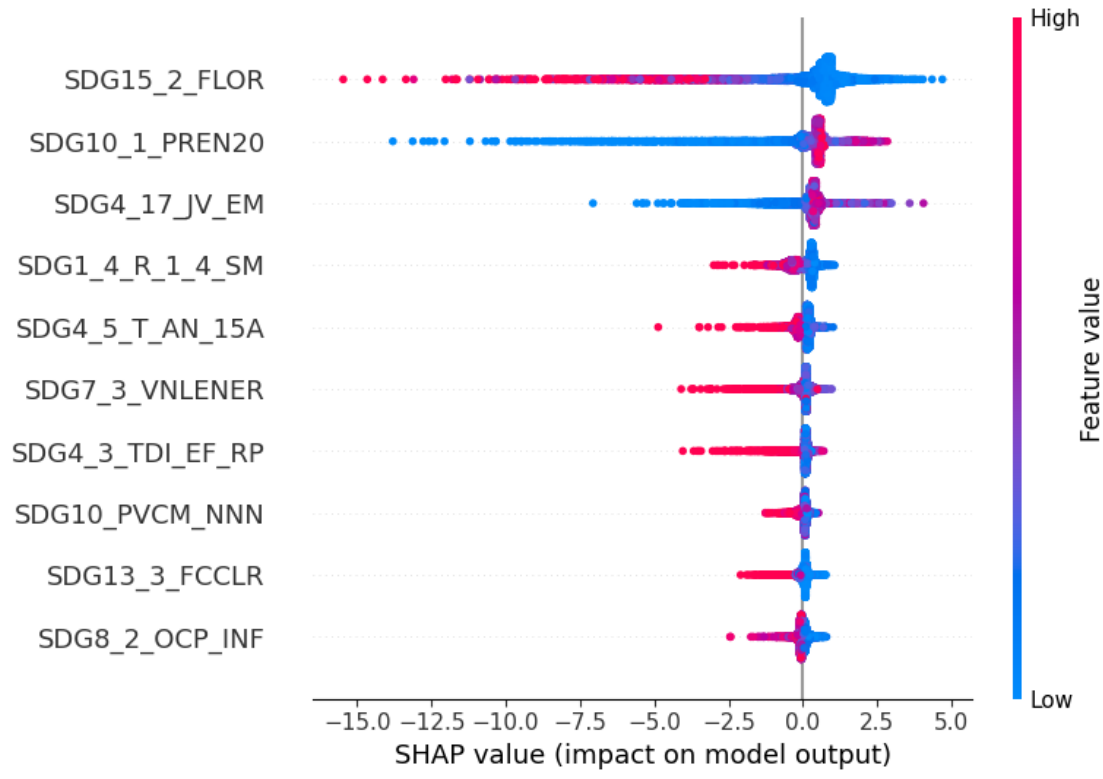


Figure 5. Top 10 SHAP values (XGBoost). Source: The authors.

Although Figure 5 indicates a negative association between the indicator of forested and natural areas per capita (SDG15_2_FLOR) and the target variable, it should be emphasized that SHAP values capture conditional statistical relationships (given the set of covariates in the model) and do not identify causality. Thus, the negative sign means that, holding the other indicators constant, higher values of forested and natural areas per capita are associated with lower values of the target variable; this does not imply that reductions in forested and natural areas lead to an increase in the percentage of households with access to electricity, or vice versa.

The indicators presented in the importance plot (Figure 4) and those shown in the SHAP plot (Figure 5) do not fully coincide because each technique addresses different questions. The importance plot was constructed from the average values of the coefficients of the selected regression model, reflecting the global contribution of each indicator according to traditional linear metrics. In contrast, SHAP values consider the marginal impact of each indicator on the model's predictions for each municipality, indicating both the magnitude and the direction of the effects.

Thus, the sets of indicators may differ because importance measures prioritize stability, whereas SHAP values exclusively reflect the internal logic of the best-performing model selected (in this case, XGBoost). This difference should be interpreted in a complementary manner: the importance ranking indicates which indicators are consistently relevant, while SHAP shows how these indicators influence the target variable—positively or negatively—at the individual (municipal) level.

To enhance the territorial interpretability of the predictive model, Figure 6 presents the geographic distribution of predicted electricity access across Brazilian municipalities. This spatial representation

doi: 10.25100/iyc.v28i1.15506

allows the identification of clusters of vulnerability, particularly in remote regions and areas with higher socioeconomic deprivation.

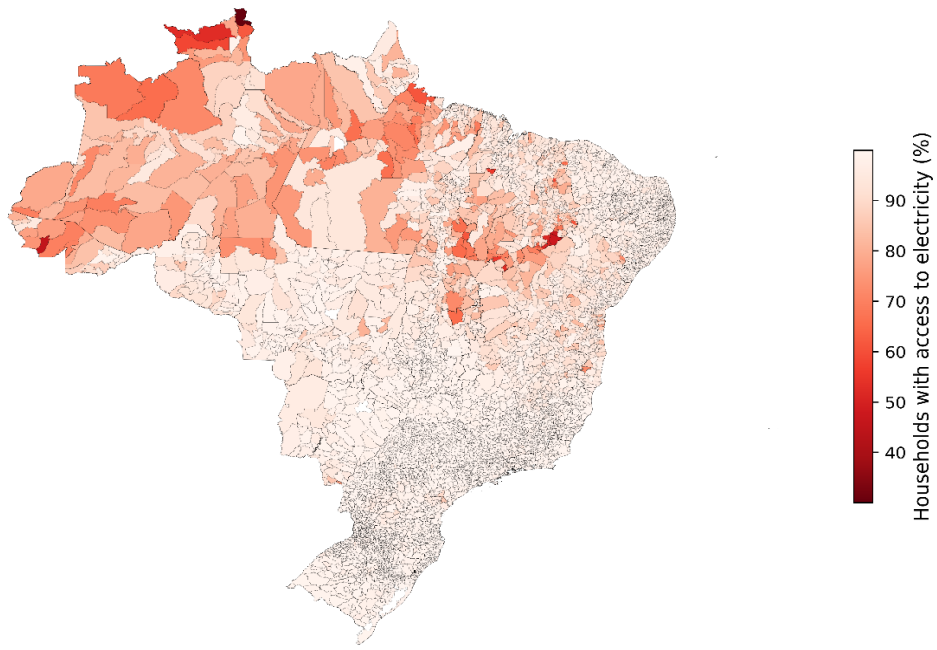


Figure 6. Spatial distribution of predicted electricity access across Brazilian municipalities.
Source: The authors.

Conclusion

This study developed a predictive model to estimate the percentage of households with access to electricity in Brazilian municipalities, based on socioeconomic variables extracted from the Sustainable Cities Development Index (SCDI). Through a robust data science pipeline, the work involved multivariate data imputation, reduction of predictor multicollinearity, selection of regression models, hyperparameter tuning, and explainability analysis using SHAP.

Among the six evaluated models—Ridge, Lasso, Random Forest, XGBoost, SVR, and MLP—XGBoost achieved the best average performance in terms of RMSE (3.42), adjusted R^2 (0.63), and MAPE (1.78%), demonstrating effectiveness in estimating the target variable. Statistical analysis confirmed significant differences among the tested models. Although Random Forest showed similar performance, the superior average predictive results of XGBoost favored its selection.

The analysis of predictor importance revealed that the most relevant indicators were income of the poorest population, forested and natural areas, and youth education. The 10 most important indicators are related to inequality and education, showing consistency with the literature on energy exclusion and structural inequalities. The SHAP analysis corroborates these findings, providing both local and global model explainability, which is particularly useful for the formulation of targeted public policies.

doi: 10.25100/iyc.v28i1.15506

The main contribution of this study is the proposal of a replicable and scalable tool for territorial diagnosis of access to electricity, based on machine learning and public secondary data. By employing SCDI indicators, the model allows the prediction of municipal electrification levels and the understanding of structural factors associated with this phenomenon. This approach represents an advance over previous, more qualitative, or localized studies.

As a limitation, the use of municipality-level aggregated data may mask intra-urban inequalities, in addition to the lack of validation with empirical field data. The study also relied on cross-sectional data, which may limit the model's ability for long-term forecasting.

Future studies may integrate geospatial data, time series, or energy consumption sensor data to increase the accuracy and granularity of the estimates. Longitudinal data may enhance the model's predictive capacity over the long term.

It is therefore concluded that the predictive modeling proposed in this work constitutes a relevant and innovative contribution to the field of electricity access analysis in Brazil, with potential applications in monitoring, prioritizing public policies, and studies on multidimensional energy poverty.

CrediT authorship contribution statement

Conceptualization - Ideas: Leandro Scala da Rocha. Data curation: Derly Henao, Merly Daza, Michell Hernández. Formal analysis: Leandro Scala da Rocha. Investigation: Leandro Scala da Rocha. Methodology: Leandro Scala da Rocha. Project Management: João Bosco Gonçalves. Resources: Leandro Scala da Rocha. Software: Leandro Scala da Rocha. Supervision: João Bosco Gonçalves. Validation: Leandro Scala da Rocha e João Bosco Gonçalves. Writing - original draft - Preparation: ALeandro Scala da Rocha e João Bosco Gonçalves. Writing - revision and editing -Preparation: Leandro Scala da Rocha e João Bosco GonçalvesFunding: not declared.

Conflict of interest: not declared. Ethical considerations: not declared.

References

1. Galindo, MF. Universal electricity access in remote areas: Building a pathway toward universalization in the Brazilian Amazon [Internet]. Stockholm (SE): KTH Industrial Engineering and Management, Royal Institute of Technology; 2014 [cited 2025 Dec 16]. Available at: 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. Last-mile rural electrification: Lessons learned from universalization programs in Brazil and Venezuela. Energy Policy. 2022; 167:113080.

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

3. Pereira JD, Santos MA, Bandeira FD, Soares FI, Vieira TA. Electrification in remote regions: An analysis of the More Light for Amazon program. Energies. 2023; 16(12):4663.

https://doi.org/10.3390/en16124663

doi: 10.25100/iyc.v28i1.15506

4. Lipscomb M, Mobarak AM, Barham T. Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil. American Economic Journal: Applied Economics. 2013; 5(2):200–231. Available at:

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

5. Brasil. Ministério de Minas e Energia (BR). Resenha energética brasileira: Exercício de 2022 [Internet]. Brasília (DF): Ministério de Minas e Energia; 2023 [cited 2025 Dec 16]. Available at: 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. Are the energy poor also income poor? Evidence from India. Energy Policy. 2012; 47:1–12.

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

7. Santillán OS, Cedano KG, Martínez M. Analysis of energy poverty in 7 Latin American countries using multidimensional energy poverty index. Energies. 2020; 13(7):1608.

https://doi.org/10.3390/en13071608

8. Wang F, Geng H, Zha D, Zhang C. Multidimensional energy poverty in China: Measurement and spatio-temporal disparities characteristics. Social Indicators Research. 2023; 168:45–78.

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

9. Freitas GF, Oliveira ML. Uma análise do Programa Luz para Todos do Governo Federal. Revista de Extensão e Estudos Rurais. 2017; 6(2):143–155.

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

10. Instituto Cidades Sustentáveis. IDSC-BR: Introdução ao Índice de Desenvolvimento Sustentável das Cidades – Brasil [Internet]. São Paulo: Instituto Cidades Sustentáveis; [date unknown] [cited 2025 Sep 15]. Available at:

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

11. Nações Unidas Brasil. Agenda 2030 para o desenvolvimento sustentável [Internet]. Brasília (DF): Nações Unidas Brasil; 2015 Sep 15 [cited 2025 Dec 16]. Available at:

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

12. Instituto Cidades Sustentáveis. IDSC-BR: Índice de Desenvolvimento Sustentável das Cidades – Brasil [Internet]. São Paulo: Instituto Cidades Sustentáveis; [date unknown] [cited 2025 Sep 15]. Available at:

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

13. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67.

doi: 10.25100/iyc.v28i1.15506

https://doi.org/10.2307/1267351

14. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B. 1996; 58(1):267–288. Available at:

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

15. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.

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

16. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]; 2016 Aug 13–17; San Francisco, CA. New York (NY): Association for Computing Machinery; 2016 [cited 2025 Dec 16]. p. 785–794).

https://doi.org/10.1145/2939672.2939785

17. Scikit-learn. sklearn.svm.SVR — scikit-learn 1.3.0 documentation [Internet]. [cited 2025 Sep 15]. Available at:

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

18. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958; 65(6):386–408. Available at:

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

19. Scikit-learn. sklearn.neural_network.MLPRegressor — scikit-learn 1.3.0 documentation [Internet]. [cited 2025 Sep 15]. Available at:

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

20. Scikit-learn. sklearn.impute.IterativeImputer — scikit-learn 1.3.0 documentation [Internet]. [cited 2025 Sep 15]. Available at:

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

21. Van Buuren S. Flexible imputation of missing data. 2nd. ed. Chapman and Hall/CRC; 2018.

22. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45(3):1–67.

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

23. Kolmogorov AN. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari. 1933; 4:83–91. Available at:

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

24. Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association. 1951; 46(253):68–78.

https://doi.org/10.1080/01621459.1951.10500769

25. Smirnov N. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics. 1948; 19(2):279–281. Available at:

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

26. Akoglu H. User's guide to correlation coefficients. Turkish Journal of Emergency Medicine. 2018; 18(3):91–93.

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

27. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. 1970; 12(3):591–612.

https://doi.org/10.1080/00401706.1970.10488699

28. Burman P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika. 1989; 76(3):503-514.

https://doi.org/10.2307/2336116

29. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika, v. 52, n. 3–4, p. 591–611, 1965.

https://doi.org/10.2307/2333709

30. Kruskal WH, Wallis W. A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 1952; 47(260):583–621.

https://doi.org/10.2307/2280779

31. Lundberg SM., Lee S-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017) [Internet]; 2017 Dec 4-9; Long Beach, CA, USA. Red Hook (NY): Curran Associates; 2017 [cited 2025 Dec 16]. p. 4765–4774. Available at:

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

32. Instituto Cidades Sustentáveis. Agenda 2030 [Internet]. São Paulo: Instituto Cidades Sustentáveis; [date unknown] [cited 2025 Sep 15]. Available at:

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

doi: 10.25100/iyc.v28i1.15506