Revista Brasileira de Estudos Regionais e Urbanos
v. 20, n. 3, p. 380-400, 2026
DOI: 1054766/rberu.v20i3.1263

Explainable Artificial Intelligence (XAI) for bounded outcomes applied to the Municipal Human Development Index

Gabriela M. Rodrigues[1] | Larissa G. Simões[2] | Gauss M. Cordeiro[3]

[1] Universidade de Sao Paulo. E-mail: gabrielar@usp.br
[2] Instituto de Pesquisa Econômica Aplicada. E-mail: larissagiardini@gmail.combr
[3] Universidade Federal de Pernambuco. E-mail: gauss@de.ufpe.br

ABSTRACT
This study analyzes the Municipal Human Development Index (MHDI) of municipalities in the state of São Paulo using unit-distribution regression models and machine learning algorithms, including Random Forest and XGBoost. The methods are compared in terms of predictive performance and interpretability. Results indicate very similar predictive accuracy across approaches, suggesting that model performance is primarily driven by the information content of the predictors and that regression models for outcomes bounded in the unit interval remain competitive with more flexible machine learning methods. Per capita household income emerged as the most influential predictor of MHDI, followed by education and poverty-related variables. Model-agnostic interpretability techniques revealed nonlinear effects and provided both local and global insights into predictor contributions. The findings offer a robust and interpretable framework for analyzing municipal development and provide evidence that may support regional monitoring and public policy planning.

KEYWORDS
Random forest, Shapley values, XGBoost

Inteligência Artificial Explicável (XAI) para resultados limitados aplicada ao Índice de Desenvolvimento Humano Municipal

RESUMO
Este estudo analisa o Índice de Desenvolvimento Humano Municipal (IDHM) de municípios do estado de São Paulo utilizando modelos de regressão para respostas no intervalo unitário e algoritmos de aprendizado de máquina, incluindo Random Forest e XGBoost. Os métodos são comparados em termos de capacidade preditiva e interpretabilidade. Os resultados indicam desempenho preditivo muito semelhante entre as abordagens, sugerindo que a capacidade dos modelos é determinada principalmente pela informação contida nos preditores e que a regressão para respostas no intervalo unitário permanece competitiva com métodos de aprendizado de máquina mais flexíveis. A renda familiar per capita emergiu como o preditor mais influente do IDHM, seguida por variáveis relacionadas à educação e à pobreza. Técnicas de interpretabilidade agnósticas ao modelo revelaram efeitos não lineares e forneceram perspectivas locais e globais sobre as contribuições dos preditores. As descobertas oferecem uma estrutura robusta e interpretável para analisar o desenvolvimento municipal e podem contribuir para o monitoramento regional e o planejamento de políticas públicas.

PALAVRAS-CHAVE
Floresta aleatória, valores Shapley, XGBoost

JEL CLASSIFICATION
R11, C55, O15

Recebido: 23/11/2025. Aceito: 15/04/2026
e-ISSN: 2447-7990 | https://revistaaber.org.br/rberu

1. Introduction

The Municipal Human Development Index (MHDI) measures the progress of a city based on three fundamental dimensions: health, education, and income. It is an important tool for assessing social progress and guiding public policies. Its prediction is a strategic tool for municipal public administrators, as it allows them to identify regions that require investment in advance, optimize efficient allocation of resources, and assess the potential impact of policies before their implementation. In addition, it contributes to public participation by highlighting regional progress and inequalities. Studies such as the Social Vulnerability Atlas (Institute for Applied Economic Research (IPEA), 2015) and the Human Development Report (United Nations Development Programme (UNDP), 2013) reinforce the relevance of MHDI to promote quality of life policies.

Several studies investigate the factors that influence MHDI in Brazilian municipalities. Silva and Gonçalves (2020) identified significant associations with electrical infrastructure, population, income, sanitation, and location. Marconato and Coelho (2019) reinforce that the regional context impacts MHDI according to local socioeconomic conditions. Meanwhile, spatial analysis of the index (UNDP; IPEA; JPF, 2013) highlights longevity, education, and income as essential components for equitable development among municipalities.

Considering that the MHDI is a continuous bounded variable with support in the interval $(0, 1)$, the use of distributions specifically designed for proportions or rates is appropriate. In this context, the beta regression model can relate the expected response variable to predictors through a link function (Ferrari and Cribari-Neto, 2004). However, to address more complex data characteristics, other distributions with restricted support have also been explored, e.g., Kumaraswamy (Kumaraswamy, 1980), log-Lindley (Gómez-Déniz et al., 2014), unit-Birnbaum-Saunders (Mazucheli et al., 2018), logit-normal and simplex distributions (Rigby et al., 2019).

Cordeiro et al. (2024) defined the odd log-logistic unit omega regression model based on quantiles to explain the MHDI and Gini index data for municipalities in São Paulo (Brazil). Rodrigues et al. (2025) analyzed the relationship between MHDI and socioeconomic variables in Rio Grande do Sul (Brazil) using an innovative odd log-logistic beta regression model. Nascimento and Gonçalves (2022) investigated the MHDI in Rio de Janeiro using Bayesian quantile regression models, highlighting the importance of considering different quantiles in the analysis.

Machine learning algorithms have gained prominence for their flexibility and high predictive performance. Random Forest (RF) (Breiman, 2001) and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) are decision tree methods (Breiman, 1984) using nonlinear and nonparametric algorithms. They do not require data distribution and automatically capture nonlinear effects and complex interactions, making them robust alternatives to traditional regression models. Despite their superior per-

formance, these models often lack comprehensibility. Its interpretation denotes the competence to comprehend and explain the inner workings of models and the factors that influence their predictions. Understanding how predictions are generated is crucial, especially in contexts such as public policymaking, where decisions need to be transparent and justifiable.

In spite of the fact that machine learning strategies have been progressively connected to financial markers, the writing still needs orderly prove on when these adaptable calculations really give focal points over conventional factual models for bounded improvement records such as the MHDI. Most existing studies focus either on parametric regression models specifically designed for proportional data or on predictive machine learning approaches, but few works compare these strategies within a unified validation framework and with a common interpretability analysis.

In this framework, the principal inquiry of the present investigation is to ascertain whether intricate machine learning algorithms substantively enhance predictive performance in comparison to distributional regression models concerning constrained socioeconomic indicators, as well as to determine if the incremental complexity yields significant advancements regarding the predictive relevance of the explanatory variables. By analyzing how different models balance predictive accuracy with variable importance, this study advances the methodological debate regarding the trade-offs between complexity, interpretability, and performance in regional economics.

Based on 2010 Brazilian census data, this study models the MHDI across municipalities in São Paulo and analyzes the impact of some socioeconomic factors. A shared validation framework is used to compare the predictive accuracy of Random Forest, XGBoost, and unit regression models. A key part of the research involves some tools to assess the relative significance of the different predictors.

2. Methodology

2.1 Data description

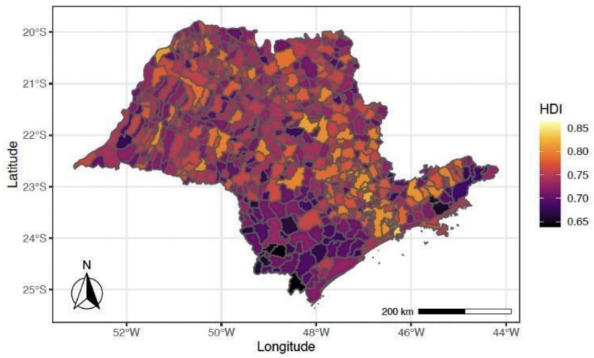
The assessment of Brazilian municipalities comprises three main dimensions:

1. Life expectancy: the average time that a person is expected to live.
2. Education: combines two subindicators to provide a comprehensive overview of access to knowledge and the quality of education in the municipality: i) Average years of schooling of the adult population (25 years or older): Reflects the level of education achieved by the adult population; and ii) Percentage of 5–6 year olds in school, plus elementary completion (11–13) and high school completion (15–17). It accesses educational progression appropriate for their age.
3. Income: the average per capita family income from the IBGE demographic census. The MHDI is equal to the geometric mean of the standardized indices of

the three aforementioned dimensions. The index ranges from 0 to 1, with higher values representing greater human development.

Data for the 645 municipalities in São Paulo are given in the Atlas of Human Development in Brazil (UNDP et al., 2013), and they are reported in Figure 1.

Figure 1. Maps of the MHDI values of cities in São Paulo from the 2010 census



Source: (Cordeiro et al., 2024)

The variables for each municipality $(i = 1, \ldots, 645)$ are reported in Table 1.

Table 1. Variables used in the analysis

Variable	Description	Unit	Source
y	MHDI	Index (0–1)	Atlas Brasil (2013)
x_1	Illiteracy rate (population aged 25 or older)	%	Atlas Brasil (2013)
x_2	Gini coefficient	Index (0–1)	Atlas Brasil (2013)
x_3	Gross higher education enrollment index	%	Atlas Brasil (2013)
x_4	proportion of children at risk of poverty.	%	Atlas Brasil (2013)
x_5	Proportion of elderly population (65+)	%	Atlas Brasil (2013)
x_6	Proportion of extremely poor population	%	Atlas Brasil (2013)
x_7	Average household income per capita (August 2010)	R$	Atlas Brasil (2013)
x_8	Proportion of people who live in homes with running water and bathrooms.	%	Atlas Brasil (2013)

Although income and education already comprise the MHDI calculation, using these variables as predictors can be useful to verify how they influence the index in real data and whether the implicit weight assigned by an institution is supported by

observed data. Modeling can also reveal which components explain the most variation of this measure among different cities, helping identify which aspects of human development are most critical in specific contexts.

2.2 Data preprocessing

2.2.1 Treatment of missing values and transformations

Prior to model estimation, the explanatory variables were preprocessed to improve numerical stability and comparability across models.

First, per capita household income (x_7) exhibited strong right-skewness. In order to decrease asymmetry and stabilize its scale, and adopt the transformed variable ($\log(x_7)$) across all model specifications.

Second, all explanatory variables were transformed to achieve a zero mean and unit variance. We standardized the test set using the mean and standard deviation computed from the training data. This step keeps test data separate to make sure the model's performance results are accurate and fair. The response variable was kept in its original scale.

2.2.2 Exploratory spatial dependence

Because the observations are municipalities within a highly integrated state, spatial dependence in the MHDI is plausible. As an exploratory diagnostic, we tested global spatial autocorrelation in y using Moran's I under randomization, with a contiguity-based spatial weights matrix (row-standardized). The null hypothesis is spatial randomness ($I = E[I]$), against the alternative of positive spatial association. The results demonstrate a significant positive spatial autocorrelation in MHDI (Moran's $I = 0.243$, $z = 9.943$, $p < 2.2 \times 10^{-16}$; Monte Carlo test with 999 permutations: $I = 0.243$, $p = 0.001$). These diagnostics suggest that municipalities with similar MHDI levels tend to cluster geographically.

2.3 Modeling approach

2.3.1 Regression models

We compared the predictive performance of several machine learning algorithms against five regression models for proportional data: Beta, Logit-Normal, and Simplex (Rigby et al., 2019), as well as two regularized linear models (Ridge and Lasso). The response variable lies in the interval $(0, 1)$, and the conditional mean μ_i is linked to the predictors via the logit function.

The probability density functions (pdfs) of the Beta, Logit-Normal (LN), and Simplex

($0 < y < 1$ and $0 < \mu < 1$) are given below.

Beta:

$$f(y) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu(1-\sigma^2)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}-1}, \tag{1}$$

where $0 < \sigma < 1$, and $\Gamma(\cdot)$ is the gamma function.

Logit Normal (LN):

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2 y(1-y)}} \exp\left(-\frac{1}{2\sigma^2}\left[\log\left(\frac{y}{1-y}\right) - \log\left(\frac{\mu}{1-\mu}\right)\right]^2\right), \tag{2}$$

where $\sigma > 0$.

Simplex:

$$f(y) = \frac{1}{(2\pi\sigma^2 y^3(1-y)^3)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\right), \tag{3}$$

where $\sigma > 0$.

All parameters are estimated by maximum likelihood. Model selection is performed on the training data using a stepwise procedure guided by the Akaike Information Criterion (AIC). The search begins from an intercept-only specification and sequentially adds or removes predictors to improve model fit.

In addition, two regularized linear models are included as benchmark specifications: Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). These models predict outcomes using a linear approach, adding a penalty (L_2/L_1) to the coefficients to prevent overfitting and handle correlated variables. We use 5-fold cross-validation on the training data to find the best regularization parameter (λ). The optimal value is chosen based on the lowest cross-validated (RMSE).

These specifications provide combined, regularized, and parametric benchmarks for comparing machine learning algorithms

2.3.2 Machine learning models

This study focuses on supervised learning for regression problems. We compare classical regression models with tree-based ensemble algorithms, which combine multiple base learners to improve overall predictive performance.

2.3.3 Decision Tree

Decision Trees (DTs), introduced by Breiman (1984), represent a supervised learning model that utilizes recursive partitioning to map out decision rules and their associated consequences. The method applies a recursive binary partitioning strategy, in which the predictor space is repeatedly divided based on the values of the explanatory variables. Each division corresponds to a *node*, and the terminal partitions are referred to as *leaves*. This study employs the minimization of the sum of squares (SQ) as the splitting criterion.

i) Let X_j be the predictor variables and define potential cut-off values s. For each pair (j, s), two regions are determined:

$$R_1(j, s) = \{X : X_j < s\}, \quad R_2(j, s) = \{X : X_j \geq s\}.$$

ii) Among all candidate pairs (j, s), the one yielding the smallest value of SQ is chosen:

$$SQ = \sum_{\{i:x_i \in R_1(j,s)\}} (y_i - \hat{y}_{R_1})^2 + \sum_{\{i:x_i \in R_2(j,s)\}} (y_i - \hat{y}_{R_2})^2.$$

iii) Splitting proceeds until a stopping point is reached, usually determined by the minimum sample size allowed in the final nodes.

A major challenge with DTs is overfitting, where the model captures noise and fits too closely to the training set, losing its ability to generalize. To alleviate this problem, pruning strategies and ensemble techniques are often employed, as they enhance stability and limit unnecessary splits.

2.3.4 Random Forest

Random Forest (RF) (Breiman, 2001) is a technique based on decision trees designed to minimize variance and bias (Hastie et al., 2009). The method generates multiple trees using bootstrap resampling and random subsets of predictor variables, and combines their outputs through aggregation. Although this approach generally improves accuracy and robustness, it is often regarded as a "black box" since the internal structure of individual trees may be difficult to interpret (Prasad et al., 2006; Rodriguez-Galiano et al., 2014).

Let (x_i, y_i) be a training dataset of size n, where x_i is a vector of p predictors for the ith response. The RF procedure can be summarized as follows.

i) Create B bootstrap samples of size from the training data. Each draw is made with replacement, so some elements may appear multiple times while others are omitted. Each sample utilizes roughly two-thirds of the observations, leaving the remaining third as the *out-of-bag (OOB)* set.

ii) For each split, choose $m < p$ predictors at random.

iii) At each node, select the best split among variables.

iv) Allow every tree to develop naturally, producing a predictor $\hat{f}^B(x)$ for Y, where $b = 1, \ldots, B$.

v) Aggregate the set of predictors to the final RF estimator:

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

2.3.5 XGBoost

The XGBoost (Chen and Guestrin, 2016) is a scalable and effective gradient boosting implementation (Friedman, 2001). It iteratively builds decision trees, each correcting the errors of the last. The final model is a weighted combination of all trees, offering a high-precision prediction.

Gradient descent is a technique for optimizing models iteratively, aiming to reach the lowest possible. A regularization term is included to avoid overfitting. Let b be the number of boosting rounds (trees) for the data size n. The algorithm sequentially adds new trees to fix the errors of the preceding one. Suppose that each tree has Z leaves and that each leaf z is assigned a weight w_z. For an observation x_i, the output of the bth tree is represented as:

$$f_b(x_i) = w_z, \quad \text{where } z \text{ is the leaf to which } x_i \text{ belongs.}$$

The model's final output is the sum of all trees' weighted contributions. The iterative strengthening process, combined with regularization, improves robustness and reduces the tendency to overfit.

2.4 Model validation and hyperparameter tuning

Model performance was evaluated using repeated train–test splits in order to account for sampling variability. The models underwent 10 iterations of Monte Carlo cross-validation, utilizing an 80/20 train-test split for training and validation in each cycle.

A summary of predictive performance was given by the mean, standard deviation, and 95% empirical intervals (2.5th–97.5th percentiles) of evaluation metrics across repeated splits.

For models requiring hyperparameter tuning (Lasso, Ridge, and XGBoost), five-fold cross-validation was employed to optimize hyperparameters from the partitioned trained data into $K = 5$ folds, which were selected based on their cross-validated performance. This approach avoids information leakage from the test set and offers an objective evaluation of out-of-sample predictive performance (Burman, 1989; Borra and Di Ciaccio, 2010).

Because the tuning procedure was repeated across the 10 train–test splits, the hyperparameter values obtained in each repetition were recorded. Final hyperparameter values were then defined using robust aggregation rules: the median was used for continuous parameters (e.g., regularization strength and number of boosting iterations), and the most frequently selected value (mode) was used for discrete parameters (e.g., tree depth and sampling rates). Final models were subsequently re-estimated on the training data using these aggregated hyperparameter values.

2.5 Performance metrics

The root mean squared error (RMSE) and coefficient of determination (R^2) determine the model performance on the held-out test set.

R^2 represents the explained variance ratio, whereas RMSE highlights the average error magnitude. Lower RMSE values and higher R^2 values indicate better out-of-sample predictive performance. They are given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{and} \quad \text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2},$$

where y_i is the observation, $\hat{y}_i$ the corresponding prediction, and $\bar{y}$ the sample mean.

2.6 Model interpretation

This study aims to predict future outcomes. Models are evaluated based on their out-of-sample performance, and interpretability tools are used to assess the contribution of predictors to predictive accuracy. Accordingly, the results identify variables that are most strongly associated with and predictive of the MHDI, rather than causal effects or policy impacts.

The distributional and regularized regression models are included as interpretable benchmark specifications. In these models, coefficient estimates are interpreted as descriptive associations conditional on the observed covariates.

2.6.1 Permutation importance

Variable importance was assessed using permutation importance (Altmann et al., 2010), a model-agnostic approach that allows direct comparison across machine learning and regression models.

For each fitted model, predictive performance was first computed on the test set. Then, for each predictor x_j, its values were randomly permuted holding all other variables constant, and predictions were recomputed. The importance of variable x_j was defined as the increase in the RMSE:

$$\Delta\text{RMSE}_j = \text{RMSE}_{\text{permuted}(j)} - \text{RMSE}_{\text{original}}.$$

Importance values were normalized and reported as the percentage contribution of each variable to the total increase in RMSE. The procedure was performed on the held-out test data, so the results reflect out-of-sample predictive relevance rather than in-sample fit.

These measures capture predictive importance and should not be interpreted as marginal effects or causal relationships. Feature importance is model-dependent and may be influenced by correlations among predictors (Molnar, 2020; Athey and Imbens, 2019).

2.6.2 SHAP values

To complement the global permutation importance results, SHAP (SHapley Additive Explanations) values were calculated for the Random Forest and XGBoost (Lundberg and Lee, 2017).

SHAP values determine the impact of each feature on a single prediction, allowing the assessment of both the direction and magnitude of variable effects. Aggregated summaries were used to obtain global interpretation measures, such as mean absolute Shapley values.

2.7 Alternative model specifications

Two modeling scenarios were considered to check result consistency, and address potential mechanical relationships between predictors and the dependent variable.

Scenario 1 (full model) includes the complete set of explanatory variables. In particular, per capita income (x_7) and education indicators (x_1 and x_3) are retained. These variables are directly related to the construction of the MHDI and they are expected to have high predictive power. After the variable selection procedure described previously, the systematic component of the distributional regression models is specified

as

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_7 x_{i7} + \beta_8 x_{i8}, \tag{4}$$

where the explanatory variables are defined in Section 2.1.

Since income and education indicators directly contribute to the calculation of the MHDI, their inclusion may introduce a mechanical or tautological relationship. To evaluate the extent to which predictive performance depends on these direct components, a reduced specification was also considered.

Scenario 2 (reduced model) excludes the variables that directly compose the MHDI, namely per capita income (x_7) and education indicators (x_1 and x_3). The systematic component reduces to

$$\text{logit}(\mu_i) = \beta_0 + \beta_2 x_{i2} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_8 x_{i8}. \tag{5}$$

This specification focuses on structural socioeconomic and demographic factors, such as inequality, poverty, housing conditions, and demographic structure, allowing the analysis to identify predictors that are not mechanically embedded in the index.

All models, distributional regressions, penalized regressions, and machine learning algorithms are estimated under both scenarios using the same preprocessing, validation, and evaluation procedures. This comparison allows the sensitivity of predictive performance and variable importance to the inclusion of the index components to be assessed.

3. Results and discussion

3.1 Descriptive statistics and dependence diagnostics

Table 2 reports some statistics of the predictor variables.

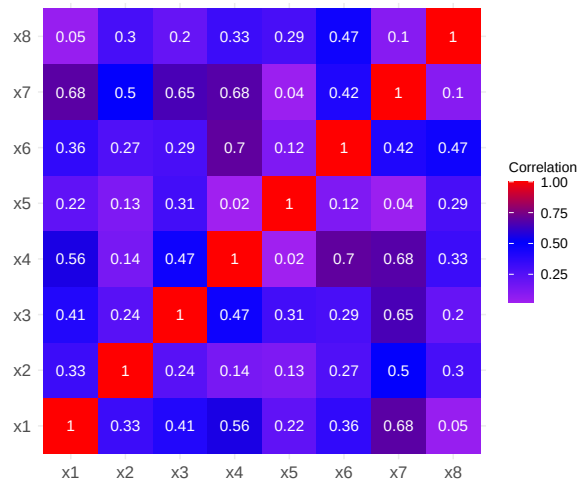
Table 2. Descriptive analysis of MHDI data

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
x_1	1.61	6.75	9.29	9.377	11.62	21.43
x_2	0.33	0.41	0.45	0.449	0.48	0.67
x_3	4.44	16.93	22.98	24.334	30.54	76.78
x_4	7.45	26.42	33.31	34.822	41.11	76.92
x_5	3.69	7.44	8.96	9.071	10.53	18.58
x_6	0.00	0.56	1.07	1.550	1.75	14.98
x_7	318.44	588.36	686.89	713.926	798.74	2043.74
x_8	75.84	97.04	98.77	97.709	99.56	100.00

Source: from the authors (2025).
x_1: Illiteracy rate ($\geq$ 25 years); x_2: Gini index; x_3: Gross education rate; x_4: % of poor children; x_5: Aging rate; x_6: % of extremely poor; x_7: Average per capita household income; x_8: % of people who live in homes with running water and bathrooms.

Pearson correlations are reported in Figure 2. Pairwise correlations do not suggest severe multicollinearity.

Figure 2. Pearson correlation matrix for the MHDI data



Source: authors (2025).

The observed spatial clustering is consistent with regional development dynamics and shared infrastructure and labor-market linkages among neighboring municipalities. Moran's I confirmed statistically significant positive spatial autocorrelation in MHDI ($I = 0.243$, $p < 0.01$). This statistic is employed here as a diagnostic tool for spatial dependence in the outcome variable and should not be interpreted as evidence of causal spillover effects.

3.2 Predictive performance: regression vs. machine learning

Predictive performance based on repeated train–test splits (10 repetitions) is summarized in Table 3. In Scenario 1 (full model), all methods exhibit very similar predictive accuracy. The mean RMSE is approximately 0.013 for all models, and average R^2 values range from 0.801 to 0.815. Tree-based ensembles (Random Forest and XGBoost) achieve slightly higher average R^2, although the empirical intervals largely overlap across methods, indicating no substantial differences in predictive accuracy. The very small standard deviations indicate high stability of predictive performance across the repeated train–test splits.

In Scenario 2 (reduced model excluding MHDI components), predictive performance decreases for all methods. The mean RMSE increases to approximately 0.017–0.018 and average R^2 values range from 0.639 to 0.682. This reduction reflects the exclusion of variables directly related to the construction of the index, confirming their strong predictive contribution. XGBoost attains the highest average R^2 in this scenario, but differences remain modest overall.

Overall, the results indicate that: (i) predictive performance is largely driven by

variables that directly compose the index, and (ii) both statistical regression models and machine learning algorithms provide comparable predictive accuracy for these data. These results suggest that model choice plays a secondary role relative to the information content of the predictors and are consistent with previous evidence that, in structured datasets, gains from complex machine learning methods may be limited relative to simpler statistical models when the predictors already capture most of the relevant information (Fernández-Delgado et al., 2014; Makridakis et al., 2018).

Table 3. Predictive performance across models based on repeated train–test splits (10 repetitions). Values report mean, standard deviation, and empirical 95% intervals for RMSE and R^2

Scenario 1: Full model						
Model	RMSE mean	RMSE sd	R^2 mean	R^2 sd	RMSE CI	R^2 CI
Lasso	0.013	0.000	0.809	0.001	[0.013, 0.013]	[0.808, 0.811]
Ridge	0.013	0.000	0.808	0.000	[0.013, 0.013]	[0.808, 0.808]
XGBoost	0.013	0.000	0.815	0.008	[0.012, 0.013]	[0.804, 0.829]
RandomForest	0.013	0.000	0.813	0.001	[0.013, 0.013]	[0.812, 0.816]
BE (Beta)	0.013	0.000	0.802	0.000	[0.013, 0.013]	[0.802, 0.802]
Simplex	0.013	0.000	0.803	0.000	[0.013, 0.013]	[0.803, 0.803]
Logit-Normal	0.013	0.000	0.801	0.000	[0.013, 0.013]	[0.801, 0.801]
Scenario 2: Reduced model (excluding MHDI components)						
Model	RMSE mean	RMSE sd	R^2 mean	R^2 sd	RMSE CI	R^2 CI
Lasso	0.017	0.000	0.655	0.000	[0.017, 0.017]	[0.655, 0.655]
Ridge	0.017	0.000	0.663	0.000	[0.017, 0.017]	[0.663, 0.663]
XGBoost	0.017	0.000	0.682	0.012	[0.016, 0.017]	[0.661, 0.701]
RandomForest	0.017	0.000	0.652	0.005	[0.017, 0.018]	[0.642, 0.657]
BE (Beta)	0.018	0.000	0.642	0.000	[0.018, 0.018]	[0.642, 0.642]
Simplex	0.018	0.000	0.642	0.000	[0.018, 0.018]	[0.642, 0.642]
Logit-Normal	0.018	0.000	0.639	0.000	[0.018, 0.018]	[0.639, 0.639]

Source: authors (2025).

3.3 Predictor relevance and explainability

The permutation importance results presented in Table 4 show a clear and consistent pattern across models and scenarios. In Scenario 1 (full model), the logarithm of per capita income (x_7 log) is by far the most influential predictor in all models, accounting for approximately 34% to 68% of the total importance. Education-related variables (x_3 and x_1) appear as the second and third most important predictors across both distributional regressions and machine learning algorithms. The remaining variables contribute only marginally to predictive performance. This result is expected, since income and education are direct components of the MHDI, indicating that much of the predictive power in the full specification is driven by variables mechanically related to the index.

Table 4. Top five predictors based on permutation importance across models

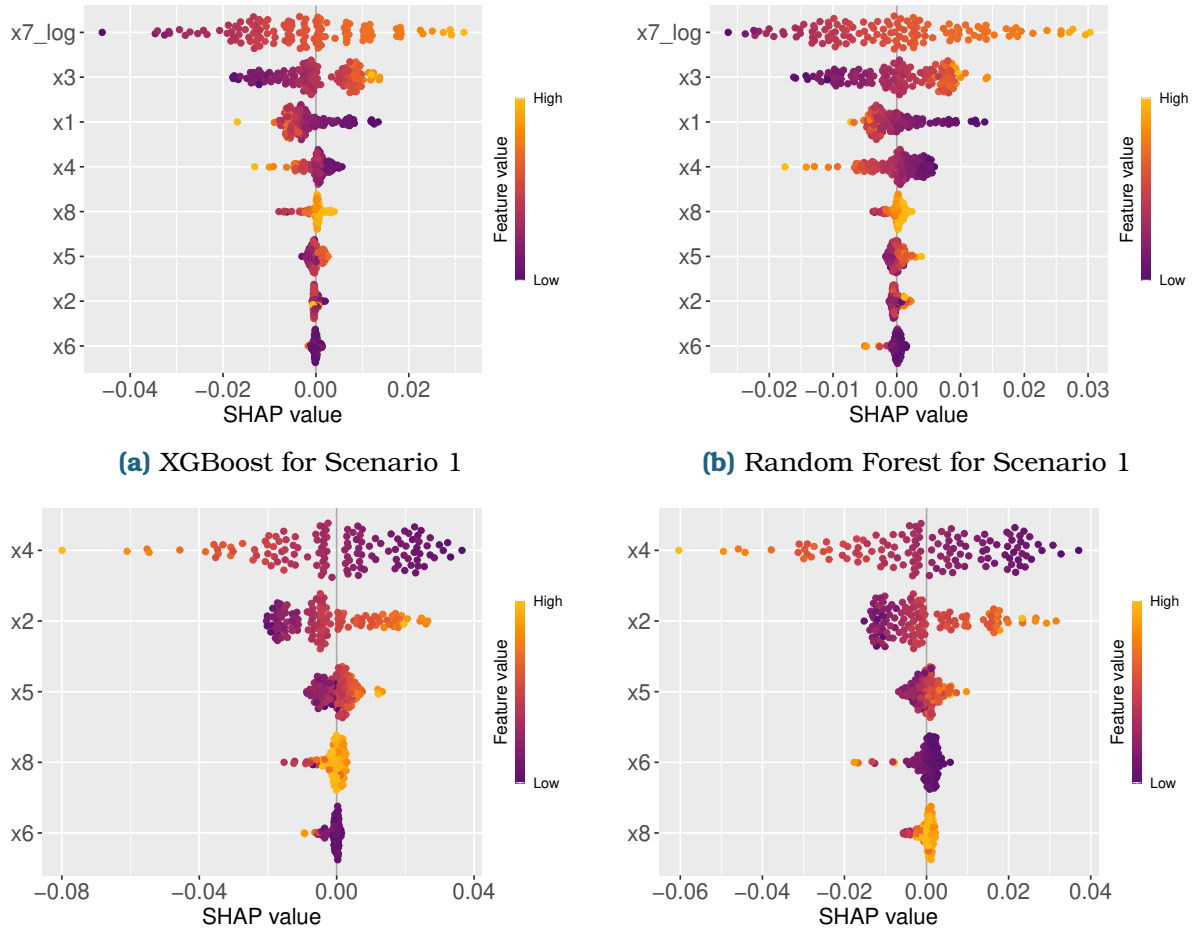
Model	Scenario 1: Full model				
Lasso	x_7 log	x_3	x_1	x_4	x_8
	54.53	22.88	15.99	5.73	0.72
Ridge	x_7 log	x_3	x_1	x_4	x_2
	34.27	26.40	19.80	16.17	1.94
RF	x_7 log	x_3	x_4	x_1	x_7
	50.73	22.41	13.23	10.58	1.53
XGBoost	x_7 log	x_3	x_1	x_4	x_5
	59.93	20.01	12.28	5.26	1.00
Beta	x_7 log	x_3	x_1	x_2	x_7
	66.24	19.23	10.54	2.46	1.01
Logit-Normal	x_7 log	x_3	x_1	x_2	x_7
	65.86	19.15	10.88	2.49	1.04
Simplex	x_7 log	x_3	x_1	x_2	x_7
	65.16	18.94	11.40	2.56	1.29
Model	Scenario 2: Reduced model (excluding MHDI components)				
Lasso	x_4	x_2	x_5	x_8	x_7
	69.51	25.66	3.89	0.48	0.46
Ridge	x_4	x_2	x_5	x_7	x_8
	65.52	27.19	4.04	2.77	0.48
RF	x_4	x_2	x_7	x_5	x_8
	65.40	23.32	5.95	4.76	0.58
XGBoost	x_4	x_2	x_5	x_7	x_8
	65.79	23.21	5.97	2.82	2.21
Beta	x_4	x_2	x_5	x_8	x_7
	69.88	25.03	4.40	0.69	0.00
Logit-Normal	x_4	x_2	x_5	x_8	x_7
	69.68	25.22	4.57	0.54	0.00
Simplex	x_4	x_2	x_5	x_8	x_7
	69.54	25.53	4.46	0.47	0.00

When these components are excluded (Scenario 2), a substantial shift in the importance structure is observed. The percentage of children vulnerable to poverty (x_4) becomes the dominant predictor in all models, explaining between 61% and 69% of the total importance. The Gini index (x_2) consistently appears as the second most relevant variable, accounting for approximately 24% to 32%. Demographic structure (x_5) and extreme poverty (x_6) show moderate contributions, while access to basic infrastructure (x_8) plays a minor role.

The ranking of predictors is highly consistent across distributional regressions (Beta, Logit-Normal, and Simplex), penalized regressions (Ridge and Lasso), and Random Forest and XGBoost methods. This agreement suggests that the identified drivers of MHDI are robust to model specification and estimation strategy.

Overall, beyond the direct components of the index, structural socioeconomic conditions, particularly child vulnerability to poverty and income inequality, emerge as the main predictors of variation in municipal human development levels.

Figure 3. SHAP summary plots for the contributions to MHDI predictors.



(a) XGBoost for Scenario 1

(b) Random Forest for Scenario 1

(c) XGBoost for Scenario 2

(d) Random Forest for Scenario 2

Source: Prepared by the authors (2025).

Figure 3 shows the SHAP plots for Random Forest and XGBoost models. The results are consistent with the permutation importance analysis. In Scenario 1, per capita income (x_7 log) shows the largest contribution to the predictions, followed by education variables (x_3 and x_1) and child poverty (x_4). Higher values of income and education are generally associated with positive contributions to predicted MHDI, while higher levels of deprivation indicators contribute negatively.

In Scenario 2, after excluding variables directly related to the construction of the index, the percentage of child poverty (x_4) becomes the main predictor, followed by income inequality (x_2). The dispersion of SHAP values indicates heterogeneous and nonlinear relationships across municipalities. These patterns are highly consistent across algorithms, and reinforce the robustness of the identified predictors. These values should be interpreted as measures of predictive contribution rather than causal effects.

3.4 Benchmark distributional model: fit and descriptive inference

After the predictive comparison, a distributional regression model is used as a benchmark to provide formal inference and diagnostic assessment under a parametric specification. Consequently, the simplex model was chosen based on AIC, Bayesian IC (BIC), and Global Deviation (GD) (Table 5). The maximum likelihood estimates (MLEs), standard errors, and p-values are reported in Table 6.

Further, the choice of the simplex model was supported by diagnostic analyses. Graphical comparisons between observed and fitted values indicated good agreement across the support of the response variable, with no systematic lack of fit in the lower or upper regions of the distribution. Residual diagnostics did not reveal patterns suggesting model misspecification or strong heteroscedasticity.

The MHDI values are concentrated away from the boundaries of the unit interval, with few observations close to 0 or 1. Under this condition, the Beta, Logit-Normal, and Simplex models exhibited similar tail behavior. The Simplex specification provided a stable fit across the entire range of the data and a parsimonious representation of the conditional mean structure.

Table 5 reports the mean values of AIC, BIC, and Global Deviance obtained across the repeated train–test splits in Scenario 1. The results indicate similar goodness-of-fit across distributions, with the Simplex model showing slightly better values.

The inferential analysis reported in this section refers to Scenario 1 (full model). After the predictive comparison based on repeated train–test splits, a final model was re-estimated using the full sample to obtain parameter estimates and residual diagnostics. This final specification is used exclusively for descriptive inference and model adequacy assessment.

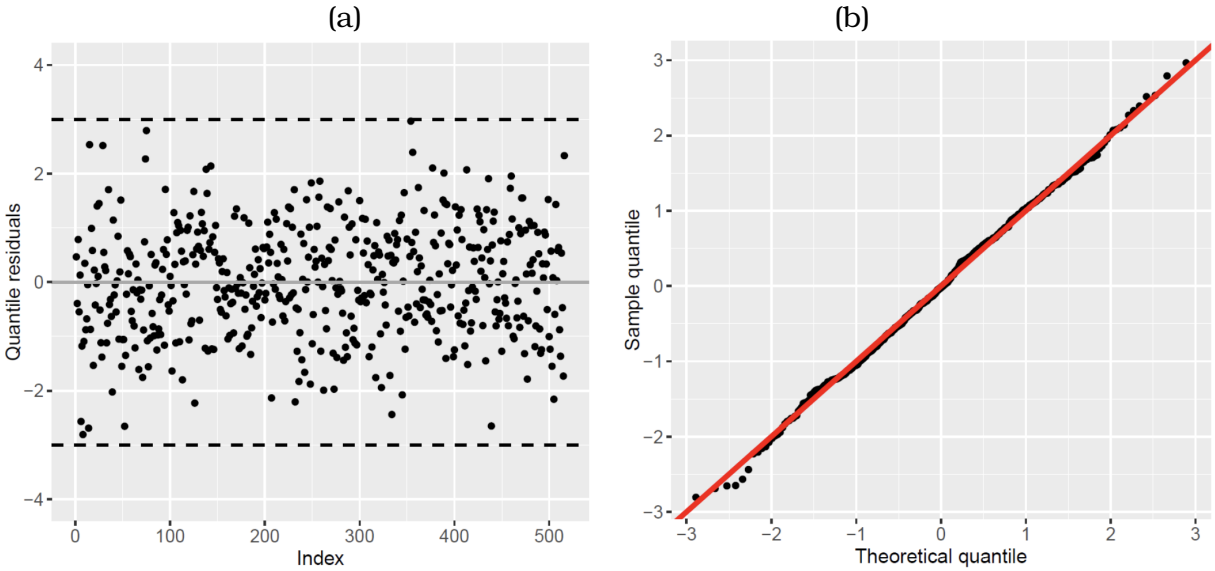
Table 5. AIC, BIC and GD of the Beta, Simplex and LN regression models

Model	AIC mean	BIC mean	GD mean
Beta	-3025.796	-2991.827	-3041.796
Logit-Normal	-3032.424	-2998.455	-3048.424
Simplex	-3036.367	-3002.398	-3052.367

Source: from the authors (2025).

Figure 4 presents the quantile residual plots for the selected Simplex model. The index plot shows residuals randomly distributed around zero and mostly within the interval $(-3, 3)$, while the normal probability plot indicates close adherence to the standard normal distribution. These findings confirm the model's validity throughout the support of the response variable.

Figure 4. Quantile residuals plots of the simplex regression model: (a) Index and (b) Normal probability plot



Source: authors (2025).

Table 6. Findings from the simplex regression analysis

Parameter	Variable	MLE	SE	p-value
β_0	(intercept)	1.050	0.003	0.000
$\beta_1(x_1)$	Illiteracy rate ($\geq$ 25 years)	-0.034	0.004	0.000
$\beta_2(x_2)$	Gini index	-0.014	0.004	0.000
$\beta_3(x_3)$	Gross tertiary enrollment rate	0.048	0.003	0.000
$\beta_6(x_6)$	Proportion of families living in extreme poverty	0.010	0.004	0.024
$\beta_7(x_7 \log)$	Average per capita household income	0.108	0.006	0.000
$\beta_8(x_8)$	Percentage of population living in households with a bathroom and running water	0.007	0.003	0.023
$\log(\sigma)$	-	-1.891	0.028	0.000

Source: authors (2025).

These results provide descriptive evidence on the relationship between the predictors and MHDI. Moran's I statistics demonstrates significant spatial autocorrelation in the response.

Because the benchmark models do not explicitly incorporate spatial-lag or spatial-error structures, this dependence calls for cautious interpretation of coefficient estimates and suggests a natural extension for future research, such as spatial econometric specifications or spatial cross-validation.

3.5 Implications and limitations

The importance measures identify the socioeconomic factors most strongly associated with out-of-sample variation in MHDI, indicating priority domains for monitoring and policy attention. Variables related to income, education, and inequality consis-

tently showed the highest predictive relevance across models, suggesting that these dimensions are closely linked to differences in municipal development levels.

In practice, the proposed framework can support public management by identifying municipalities at higher risk of low human development and guiding the prioritization of monitoring and resource allocation for diagnostic and planning purposes.

In addition, the analysis is based on cross-sectional data, which limits temporal inference and the evaluation of dynamic changes in municipal development.

4. Final considerations

This study examined the correlation between Municipal Human Development Index (HDI) scores within the state of São Paulo and a set of socioeconomic variables by comparing machine learning algorithms with unit-distribution regression models. The results showed very similar predictive performance across methods, indicating that model choice plays a secondary role relative to the information content of the predictors. These findings also demonstrate that regression models specifically designed for outcomes bounded in the unit interval remain competitive with more flexible machine learning approaches.

Among the evaluated methods, Random Forest achieved the highest average predictive performance, although differences across models were modest. The most relevant predictors were per capita household income, tertiary education enrollment, illiteracy rate (population aged $\geq$ 25 years), and the percentage of children vulnerable to poverty. Municipalities with higher income and education levels and lower deprivation indicators tended to present higher MHDI values. The distributional regression models identified a similar set of relevant variables, with minor differences in the inclusion of extreme poverty indicators.

Shapley Additive Explanations (SHAP) made machine learning models interpretable by revealing nonlinear effects and both local and global patterns. In contrast, regression models provide a parametric framework that allows statistical inference, including hypothesis testing and parameter interpretation. Together, these approaches offer complementary insights for predictive analysis and substantive interpretation.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725.

Borra, S. and Di Ciaccio, A. (2010). Measuring the prediction error. a comparison of

cross validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, 54(12):2976–2989.

Breiman, L. (1984). *Classification and Regression Trees*. Routledge, New York, NY, USA, 1 edition.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA.

Cordeiro, G. M., Rodrigues, G. M., Prataviera, F., and Ortega, E. M. (2024). A new quantile regression model with application to human development index. *Computational Statistics*, 39(6):2925–2948.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Gómez-Déniz, E., Sordo, M. A., and Calderín-Ojeda, E. (2014). The log–lindley distribution as an alternative to the beta regression model with applications in insurance. *Insurance: Mathematics and Economics*, 54(1):49–57.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. Springer-Verlag, New York, NY, USA, 2 edition.

Institute for Applied Economic Research (IPEA) (2015). Atlas da vulnerabilidade social nos municípios brasileiros. https://repositorio.ipea.gov.br/bitstream/11058/4381/1/Atlas_da_vulnerabilidade_social_nos_ Accessed: 18 April 2025.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2):79–88.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, Long Beach, CA, USA.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889.

Marconato, M. and Coelho, M. H. (2019). O idhm dos municípios brasileiros sob a perspectiva da análise exploratória de dados espaciais. *Economia & Região*, 7(2):49–66.

Mazucheli, J., Menezes, A. F., and Dey, S. (2018). The unit-birnbaum-saunders distribution with applications. *Chilean Journal of Statistics*, 9(1):47–57.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com. Online book available at https://christophm.github.io/interpretable-ml-book/.

Nascimento, M. G. and Gonçalves, K. C. (2022). Bayesian variable selection in quantile regression with random effects: an application to municipal human development index. *Journal of Applied Statistics*, 49(13):3436–3450.

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Chapman and Hall/CRC, Boca Raton, FL, USA, 1 edition.

Rodrigues, G. M., Cordeiro, G. M., Ortega, E. M., and Vila, R. (2025). New regression model and machine learning for fitting proportional data with application. *Statistics*, 59(2):498–516.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., and Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (southern spain). *Science of the Total Environment*, 476:189–206.

Silva, C. A. d. and Gonçalves, R. F. (2020). Fatores que impactam o idh em municípios brasileiros. *Revista de Administração, Contabilidade e Economia*, 19(1):1–20.

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

UNDP, IPEA, and Fundação João Pinheiro (2013). Atlas of human development in brazil. http://www.atlasbrasil.org.br. Accessed: 2026-02-15.

United Nations Development Programme (UNDP) (2013). Institute for applied economic research (ipea) and joão pinheiro foundation (jpf), atlas do desenvolvimento humano no brasil 2013.

Conflicts of interest and the use of artificial intelligence

The authors declare no competing interests in this work.

Artificial intelligence tools serve solely to assist with data analysis and interpretation. In particular, we adopted the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms to model the data. We also applied SHAP methods to enhance interpretability and ensure transparency of the results. These approaches were used as complementary resources to strengthen the robustness and clarity of the findings, without influencing the scientific conclusions drawn by the authors.