Cadernos do Desenvolvimento Fluminense - ISSN: 2317-6539
Rio de Janeiro | n. 31 | jan./jun. – 2026
Rio de Saberes: estudo sobre o território fluminense
Dossiê: II Seminário de Estudos sobre o Estado do Rio de Janeiro (II SEERJ)
ARTIGO

DESENVOLVIMENTO SUSTENTÁVEL MUNICIPAL NO ESTADO DO RIO DE JANEIRO: UMA ANÁLISE MULTIVARIADA BASEADA EM CIÊNCIA DE DADOS

SUSTAINABLE MUNICIPAL DEVELOPMENT IN THE STATE OF RIO DE JANEIRO: A MULTIVARIATE ANALYSIS BASED ON DATA SCIENCE

Leandro Scala da Rocha
https://orcid.org/0009-0002-9920-7337
Correspondência: scala.leandro@ufrrj.br
Universidade Candido Mendes, Campos dos Goytacazes, Rio de Janeiro, Brasil.

Geísa Pereira Marcilio Nogueira
https://orcid.org/0000-0002-2122-2935
Correspondência:
Universidade Candido Mendes, Campos dos Goytacazes, Rio de Janeiro, Brasil.

Lia Hasenclever
https://orcid.org/0000-0003-1384-6323
Correspondência: lia.hasenclever@ucam-campos.br
Universidade Candido Mendes, Campos dos Goytacazes, Rio de Janeiro, Brasil.

DOI: 10.12957/cdf.2026.96978
Recebido em: 21 fev. 2026 | **Aceito em:** 13 maio 2026

RESUMO

Este estudo analisou o Índice de Desenvolvimento Sustentável das Cidades (IDSC) dos municípios do estado do Rio de Janeiro (ERJ), com o objetivo de identificar os principais fatores associados ao seu desempenho. Foram utilizados os indicadores oficiais do IDSC, submetidos a tratamento de dados ausentes por imputação multivariada, redução de multicolinearidade por correlação e *Variance Inflation Factor* (VIF) e modelagem por algoritmos de *Machine Learning* (Lasso, Ridge e *Random Forest*). A seleção do modelo foi realizada por validação cruzada, sendo o Ridge o de melhor desempenho. As importâncias dos indicadores foram avaliadas por meio dos coeficientes do modelo e da análise de explicabilidade com valores SHAP (*SHapley Additive exPlanations*). Os resultados evidenciaram que o percentual de esgoto tratado antes de chegar ao mar, rios e córregos, o índice de tratamento de esgoto e o percentual da população atendida com esgotamento sanitário foram os principais indicadores explicativos do desempenho sustentável dos municípios do ERJ. Diferentemente de estudos anteriores, predominantemente descritivos, este trabalho aplicou uma abordagem quantitativa integrada para identificar padrões estruturais nos dados. Como contribuição, o estudo fornece evidências empíricas que podem subsidiar a priorização

de políticas públicas locais alinhadas à Agenda 2030, especialmente em áreas de infraestrutura sanitária e sustentabilidade urbana.
Palavras-chave: desenvolvimento regional; planejamento regional; indicadores socioeconômicos; sustentabilidade urbana.

ABSTRACT

This study analyzed the Sustainable Cities Development Index (SCDI) of municipalities in the state of Rio de Janeiro (SRJ), aiming to identify the main factors associated with its performance. Official SCDI indicators were used, undergoing missing data treatment through multivariate imputation, multicollinearity reduction via correlation analysis and Variance Inflation Factor (VIF) and modeling using Machine Learning algorithms (Lasso, Ridge, and Random Forest). Model selection was carried out through cross-validation, with Ridge achieving the best performance. Indicator importance was assessed using model coefficients and explainability analysis based on SHAP values (SHapley Additive exPlanations). The results showed that the percentage of sewage treated before reaching the sea, rivers, and streams, the sewage treatment index, and the percentage of the population served by sanitation systems were the most relevant explanatory indicators of sustainable performance in SRJ municipalities. Unlike previous studies, which were predominantly descriptive, this work applied an integrated quantitative approach to identify structural patterns in the data. As a contribution, the study provides empirical evidence that can support the prioritization of local public policies aligned with the 2030 Agenda, particularly in areas related to sanitation infrastructure and urban sustainability.
Keywords: regional development; regional planning; socioeconomic indicators; urban sustainability.

1 INTRODUÇÃO

O Índice de Desenvolvimento Sustentável das Cidades (IDSC) foi desenvolvido pelo Instituto Cidades Sustentáveis, como ferramenta para avaliar o progresso municipal em direção aos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 da Organização das Nações Unidas (Instituto Cidades Sustentáveis, s.d.-a; Nações Unidas Brasil, 2015).

O IDSC traduz as metas globais dos ODS em indicadores aplicáveis ao contexto local, possibilitando que gestores públicos e a sociedade civil acompanhem, de maneira objetiva e comparável, os avanços e desafios associados à sustentabilidade urbana. O índice assume valores no intervalo entre 0 e 100, significando, portanto, a porcentagem do desempenho ótimo que um município atingiu.

Os 17 ODS a serem alcançados até o ano 2030 foram estabelecidos em 2015, em comum acordo por 193 países, incluindo o Brasil (Instituto Cidades Sustentáveis, s.d.-a).

Identificar os fatores associados ao IDSC é importante para compreender quais fatores mais contribuem para o desempenho municipal e, assim, orientar políticas públicas mais eficazes e baseadas em evidências (Pinheiro, 2022). Essa identificação contribui para alocar recursos de forma estratégica, priorizar áreas críticas e promover sinergias entre diferentes dimensões do desenvolvimento sustentável.

No caso do ERJ, marcado por fortes desigualdades socioeconômicas e desafios urbanos estruturais, a análise dos fatores associados ao IDSC ganha importância (Instituto Jones dos Santos Neves, 2025). Estudos regionais permitem considerar especificidades locais, contribuindo para políticas públicas mais contextualizadas e efetivas (Guimarães, 1997).

Pesquisas recentes vêm explorando o uso do IDSC em diferentes contextos. Wissmann e Backes (2022) analisaram o índice sob uma perspectiva descritiva, comparando *scores* entre municípios e regiões brasileiras, sem recorrer à modelagem estatística para identificar fatores associados. De forma semelhante, Costa e Fernández (2024) utilizaram o IDSC para mapear a sustentabilidade dos municípios do estado do Paraná, detalhando o desempenho em cada ODS, mas também sem aplicar métodos preditivos ou de explicabilidade.

Este artigo propõe preencher uma lacuna, ao desenvolver um modelo de *Machine Learning* (ML) para a predição do IDSC dos municípios do ERJ. O diferencial desse trabalho é a aplicação integrada de um *pipeline* de ciência de dados contemplando desde a limpeza e normalização dos dados, o tratamento de dados ausentes e multicolinearidade de indicadores, até a estimação de modelos via validação cruzada e a seleção do melhor modelo. A abordagem ainda compreende a identificação das importâncias dos indicadores e a análise da capacidade de explicar os resultados do modelo por meio de valores SHAP, garantindo rigor estatístico e explicabilidade.

2 OBJETIVOS

O objetivo principal deste artigo é identificar os indicadores com maior poder explicativo na predição do IDSC do ERJ, dentre os indicadores municipais que o compõem, através dos seguintes objetivos específicos: (1) Estimar um modelo de ML para predição do IDSC; (2) identificar os indicadores de maior importância para a

predição do IDSC e (3) identificar o sentido e magnitude da contribuição dos indicadores na predição do IDSC.

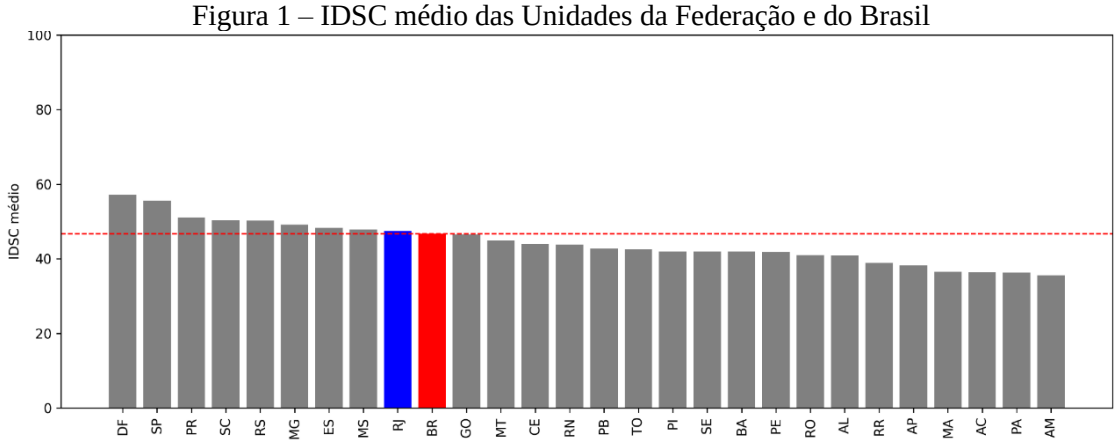
3 METODOLOGIA

Foi realizada uma pesquisa quantitativa e aplicada, de caráter descritivo-analítico com modelagem preditiva baseada em ML, a partir dos dados do IDSC do ano 2024 (IDSC-BR_2024). Esses dados são de acesso público e gratuito, compreendendo 100 indicadores para 5.570 municípios do Brasil (Instituto Cidades Sustentáveis, s.d.-b). O trabalho foi realizado em cinco etapas: (1) Seleção dos dados do ERJ; (2) tratamento de dados ausentes; (3) tratamento de indicadores colineares; (4) estimação e seleção do melhor modelo de ML e (5) análise de importâncias e de valores SHAP do melhor modelo.

3.1 Seleção dos dados do ERJ

O estudo utilizou os dados dos 100 indicadores que compõem o IDSC para todos os 92 municípios do ERJ. O IDSC médio do ERJ é 47,45 e do Brasil é 46,69, conforme Figura 1.

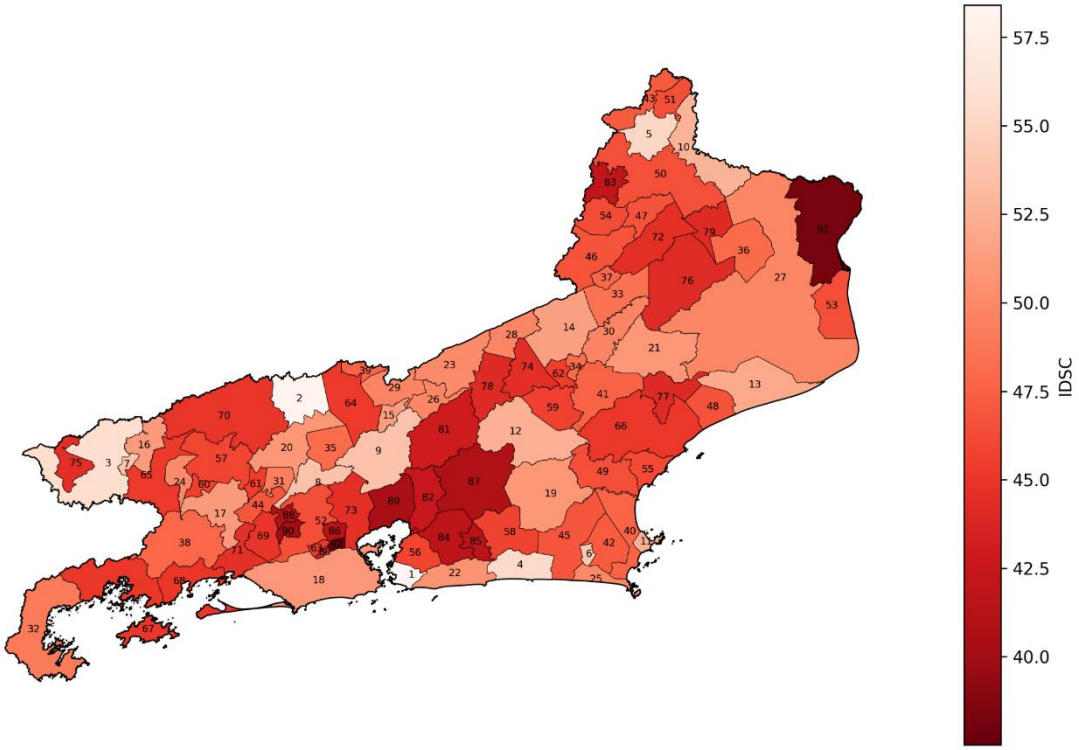
Figura 1 – IDSC médio das Unidades da Federação e do Brasil



Fonte: Elaboração própria (2025).

A Figura 2 mostra a distribuição espacial, a escala de valores e o *ranking* do IDSC dos municípios do ERJ.

Figura 2 – IDSC dos municípios do ERJ



Legenda: 1-Niterói, 2-Rio das Flores, 3-Resende, 4-Saquarema, 5-Natividade, 6-Iguaba Grande, 7-Porto Real, 8-Miguel Pereira, 9-Petrópolis, 10-Bom Jesus do Itabapoana, 11-Armação dos Búzios, 12-Nova Friburgo, 13-Quissamã, 14-Cantagalo, 15-Areal, 16-Quatis, 17-Piraí, 18-Rio de Janeiro, 19-Silva Jardim, 20-Vassouras, 21-Santa Maria Madalena, 22-Maricá, 23-Sapucaia, 24-Volta Redonda, 25-Arraial do Cabo, 26-São José do Vale do Rio Preto, 27-Campos dos Goytacazes, 28-Carmo, 29-Três Rios, 30-São Sebastião do Alto, 31-Engenheiro Paulo de Frontin, 32-Paraty, 33-Itaocara, 34-Macuco, 35-Paty do Alferes, 36-Cardoso Moreira, 37-Aperibé, 38-Rio Claro, 39-Comendador Levy Gasparian, 40-Cabo Frio, 41-Trajano de Moraes, 42-São Pedro da Aldeia, 43-Porciúncula, 44-Paracambi, 45-Araruama, 46-Santo Antônio de Pádua, 47-São José de Ubá, 48-Carapebus, 49-Casimiro de Abreu, 50-Itaperuna, 51-Varre-Sai, 52-Nova Iguaçu, 53-São João da Barra, 54-Miracema, 55-Rio das Ostras, 56-São Gonçalo, 57-Barra do Piraí, 58-Rio Bonito, 59-Bom Jardim, 60-Pinheiral, 61-Mendes, 62-Cordeiro, 63-Mesquita, 64-Paraíba do Sul, 65-Barra Mansa, 66-Macaé, 67-Angra dos Reis, 68-Mangaratiba, 69-Seropédica, 70-Valença, 71-Itaguaí, 72-Cambuci, 73-Duque de Caxias, 74-Duas Barras, 75-Itatiaia, 76-São Fidélis, 77-Conceição de Macabu, 78-Sumidouro, 79-Italva, 80-Nilópolis, 81-Teresópolis, 82-Guapimirim, 83-Laje do Muriaé, 84-Itaboraí, 85-Tanguá, 86-Belford Roxo, 87-Cachoeiras de Macacu, 88-Japeri, 89-Magé, 90-Queimados, 91-São Francisco de Itabapoana, 92-São João de Meriti.
Fonte: Elaboração própria (2025).

3.2 Tratamento de dados ausentes

Alguns indicadores componentes do IDSC do ERJ apresentaram valores ausentes para alguns municípios. Dados ausentes devem ser tratados, seja por imputação de valores ou remoção do indicador da base de dados utilizada, para evitar viés e perda de desempenho, garantindo integridade estatística e melhor generalização dos modelos de ML (Little; Rubin, 1987).

Neste trabalho foi realizada a imputação dos valores ausentes por meio do método *Multivariate Imputation by Chained Equations* (MICE). Esse método ajusta, de forma

iterativa, modelos de regressão para cada indicador com valores ausentes, utilizando os demais indicadores previamente normalizados como preditores, até a convergência dos valores imputados. A escolha deve-se à sua capacidade de preservar relações multivariadas, evitando vieses introduzidos por imputações simplistas, como médias ou medianas (Scikit-Learn, s.d.; Buuren, 2018; Buuren; Groothuis-Oudshoorn, 2011).

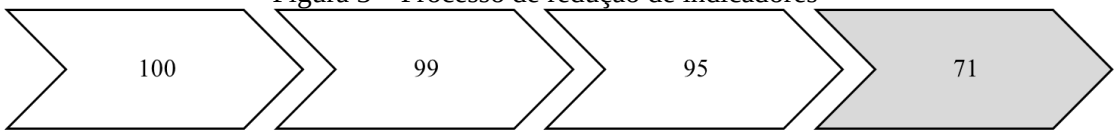
Para avaliar a plausibilidade das imputações, aplicou-se o teste de Kolmogorov–Smirnov, comparando a distribuição dos valores observados e imputados de cada indicador. Os indicadores cuja distância D foi superior a $D_{crítica}$, com significância estatística de 5% ($p < 0,05$), foram removidos por indicarem imputações inconsistentes (Kolmogorov, 1933; Massey, 1951; Smirnov, 1948).

3.3 Tratamento de indicadores colineares

Os indicadores colineares devem ser tratados para evitar multicolinearidade, que compromete a estabilidade e interpretação dos modelos de ML. A base de dados deve ser reduzida ao menor tamanho possível sem perda informacional, garantindo eficiência ao reduzir a complexidade computacional e robustez analítica ao eliminar ruídos.

O tratamento de indicadores colineares foi realizado em duas etapas: (1) foram removidos indicadores redundantes em pares com correlação linear absoluta forte ($|r| > 0,9$) (Akoglu, 2018; Pearson, 1896), preservando-se sempre aquele com maior correlação com a variável-alvo ─ IDSC; (2) aplicou-se a análise do *Variance Inflation Factor* (VIF) removendo iterativamente os indicadores até que todos apresentassem $VIF < 10$ (Marquardt, 1970). A Figura 3 mostra o processo de redução de indicadores até 71.

Figura 3 – Processo de redução de indicadores



-1 ($D > D_{crítica}$, $p < 0,05$) - 4 ($|r| > 0,9$) -24 ($VIF > 10$)

Fonte: Elaboração própria (2025).

3.4 Estimação e seleção do melhor modelo de ML

Foram estimados três modelos de ML: Regressão Ridge (Hoerl; Kennard, 1970), Regressão Lasso (Tibshirani, 1996) e *Random Forest* (Breiman, 2001). A seleção do modelo mais adequado baseou-se no maior valor médio do coeficiente de determinação

(R^2) obtido por validação cruzada do tipo *k-fold* (k = 5) (Burman, 1989), repetida 150 vezes. Esse número de repetições (B = 150) foi considerado suficiente para que o erro padrão da média do R^2 fosse inferior a 10% do valor médio com 95% de confiança (Yang *et al*., 2011), garantindo representatividade e robustez aos resultados. Além disso, a estabilidade dos coeficientes do modelo selecionado foi avaliada pela variabilidade de seus valores ao longo dos *folds* (Hastie; Tibshirani; Friedman, 2009; James *et al*., 2013; Kuhn; Johnson, 2013).

3.5 Análise de importâncias e de valores SHAP

Os dez indicadores de maior importância foram identificados a partir dos coeficientes padronizados do melhor modelo selecionado. Adicionalmente, foi realizada a análise de explicabilidade por meio do método SHAP (Lundberg; Lee, 2017), aplicado ao modelo final. Esse método fornece tanto o sentido quanto a magnitude da contribuição de cada indicador na predição do IDSC, permitindo uma explicação transparente dos resultados.

4 RESULTADOS E DISCUSSÃO

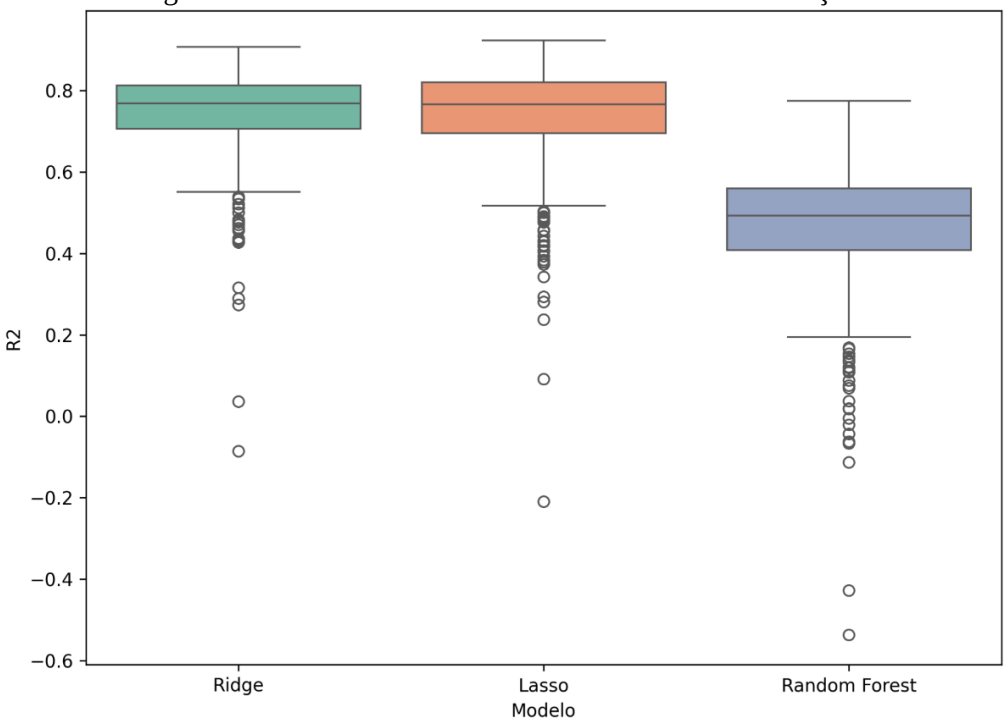
4.1 Modelo de ML

A Figura 4 apresenta os valores R^2 obtidos a partir das 750 simulações ($k \times B$) de cada modelo. Nota-se que o Ridge apresenta a maior mediana, seguido do Lasso e *Random Forest*. O teste de Shapiro–Wilk (Shapiro; Wilk, 1965) indicou violação da hipótese de normalidade para os valores de R^2 dos três modelos analisados ($p < 0,05$), ao nível de significância de 5%. Assim, foram aplicados testes não paramétricos nas análises subsequentes. O teste de Kruskal–Wallis (Kruskal; Wallis, 1952) sugeriu diferença entre os modelos avaliados. A comparação pareada utilizando o teste de Dunn, aplicado *post hoc*, com correção de Bonferroni (Dunn, 1964), indicou que não há diferença entre Ridge e Lasso, porém esses diferiram do *Random Forest*.

Todavia, cabe ressaltar que os R^2 empregados nas análises foram estimados a partir de um procedimento de validação cruzada com repetições. Esse método implica reutilização dos mesmos municípios em diferentes subconjuntos de treino e validação dos modelos avaliados gerando, portanto, dependência entre as observações.

Nesse contexto, os resultados dos testes estatísticos não devem ser interpretados sob a ótica inferencial clássica. Em vez disso, assumem caráter exploratório, servindo como instrumento para comparação entre modelos. Assim, a evidência mais robusta decorre da análise das distribuições de desempenho e de suas estatísticas descritivas, e não da significância estatística tradicional.

Figura 4 – Variabilidade dos valores de R^2 em 750 simulações



Fonte: Elaboração própria (2025).

A Tabela 1 apresenta as métricas de desempenho dos modelos avaliados. Observa-se que o Ridge apresentou o maior valor médio do R^2 (0,7488), superando os modelos Lasso (0,7415) e *Random Forest* (0,4699). Esse resultado confirma a adequação do Ridge frente às características do conjunto de indicadores do IDSC dos municípios do ERJ e, portanto, foi o modelo selecionado para as análises subsequentes.

Tabela 1 – Métricas de desempenho dos modelos

Modelo	R^2 (média)	R^2 (desvio-padrão)
Ridge	0,7488	0,0986
Lasso	0,7415	0,1136
Random Forest	0,4699	0,1352

Fonte: Elaboração própria (2025).

Considerando que a redução do número de indicadores descrita nas subseções 3.2 e 3.3 foi conduzida com base em todo o conjunto de dados — visando capturar a estrutura geral de redundância entre os indicadores do IDSC —, os resultados obtidos pelos modelos devem ser compreendidos como uma medida de desempenho condicionada a um espaço de indicadores previamente definido. Dessa forma, não se trata de uma validação preditiva rigorosa em um contexto totalmente fora da amostra, mas sim de uma análise da capacidade explicativa e preditiva dos modelos dentro de uma estrutura de dados já organizada.

4.2 Indicadores de maior importância

As importâncias dos indicadores foram estimadas a partir dos coeficientes absolutos ($|\beta|$) do Ridge, que quantificam a contribuição individual de cada indicador para a predição do IDSC. Os dez indicadores com os maiores valores absolutos de β são os de maior importância relativa e são mostrados em ordem decrescente no Quadro 1.

Quadro 1 – Descrição dos dez indicadores de maior importância

(**Continua**)

Código	**ODS**	**Indicador**
SDG14_1_ESGT	14 – Vida na Água	Esgoto tratado antes de chegar ao mar, rios e córregos (%)
SDG6_7_ESG_SAN	6 – Água Potável e Saneamento	População atendida com esgotamento sanitário (%)
SDG6_8_CLT_DML	6 – Água Potável e Saneamento	Índice de tratamento de esgoto (%)
SDG16_3_M_ARM_FG	16 – Paz, Justiça e Instituições Eficazes	Mortes por armas de fogo (100 mil habitantes)
SDG15_8_UNI_CNS	15 – Vida Terrestre	Unidades de conservação de proteção integral e uso sustentável (%)
SDG11_5_DMC_FVL	11 – Cidades e Comunidades Sustentáveis	Domicílios em favelas (%)
SDG2_1_OBS_INF	2 – Fome Zero e Agricultura Sustentável	Obesidade infantil (%)
SDG17_1_INVST_PB	17 – Parcerias e Meios de Implementação	Investimento público (R$ *per capita*)

Código	ODS	Indicador
SDG9_2_EMP_INT	9 – Indústria, Inovação e Infraestrutura	Participação dos empregos formais em atividades intensivas em conhecimento e tecnologia (%)
SDG4_18_D_SUP_EI	4 – Educação de Qualidade	Professores com formação em nível superior - Educação Infantil - rede pública (%)

Fonte: Elaboração própria (2025).

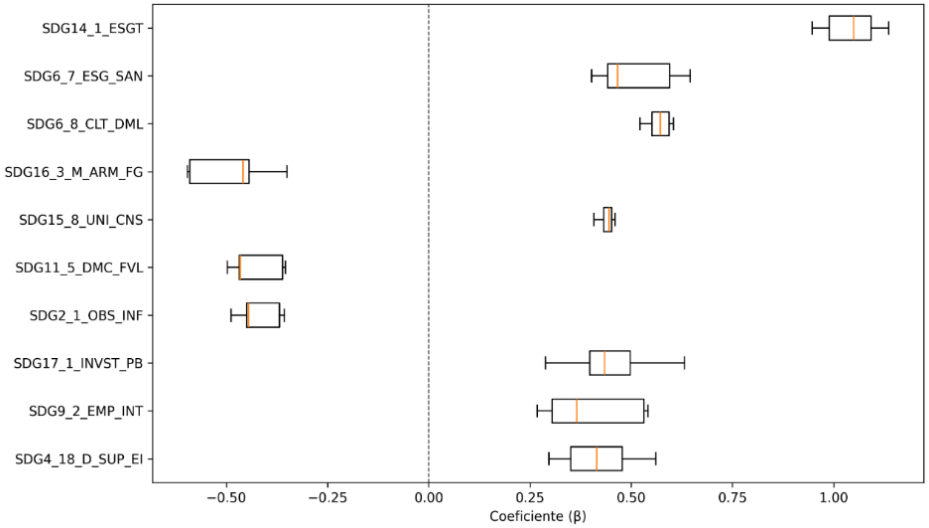
O indicador de maior importância mostrado no Quadro 1 se relaciona com Vida na Água (ODS-14), seguido de dois indicadores relacionados à água potável e saneamento (ODS-6). Porém, nota-se que esses três indicadores estão relacionados ao esgotamento sanitário. A Figura 5 mostra todos os 17 ODS da Agenda 2030.

Figura 5 – Objetivos de Desenvolvimento Sustentável (ODS)

Fonte: Nações Unidas Brasil (2015).

A Figura 6 mostra a variabilidade dos dez coeficientes β mais importantes do Ridge, obtidos a partir da validação em cinco *folds*. Nota-se a consistência nos sinais e magnitudes dos coeficientes, indicando estabilidade estatística e conferindo confiança à interpretação dos resultados.

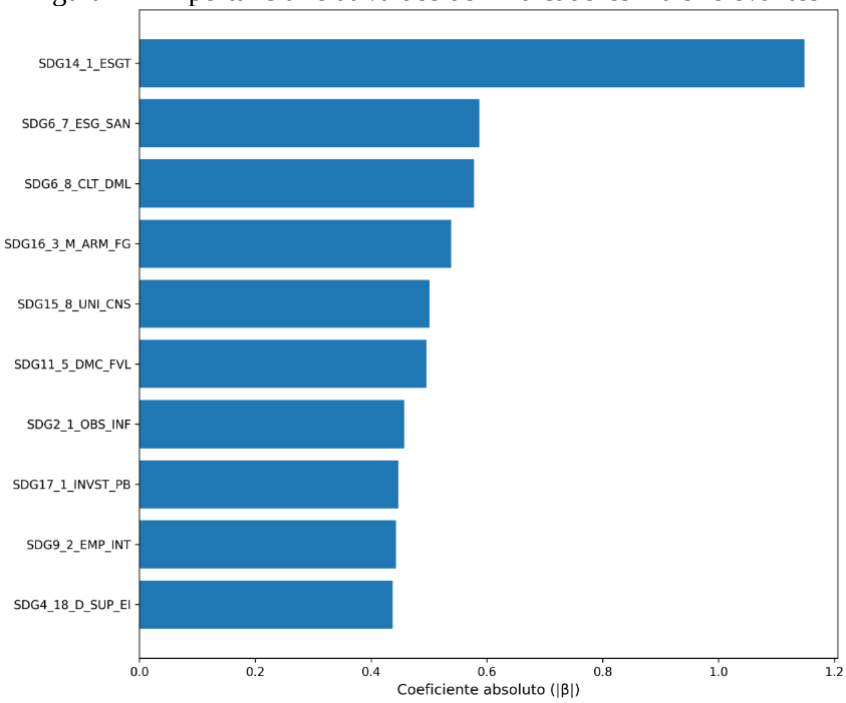
Figura 6 – Variabilidade dos dez coeficientes mais importantes do modelo Ridge



Fonte: Elaboração própria (2025).

A Figura 7 apresenta as importâncias relativas dos dez principais indicadores do IDSC. Quanto maior o valor $|\beta|$, mais importante é o indicador para a predição do IDSC. Destacam-se o percentual de esgoto tratado antes de chegar ao mar, rios e córregos (SDG14_1_ESGT), o percentual da população atendida com esgotamento sanitário (SDG6_7_ESG_SAN) e o índice de tratamento de esgoto (SDG6_8_CLT_DML), refletindo a relevância do saneamento para a sustentabilidade municipal. Indicadores sociais, como mortes por armas de fogo (SDG16_3_M_ARM_FG) e de saúde, como o percentual de obesidade infantil (SDG2_1_OBS_INF), também aparecem entre os mais relevantes, evidenciando a associação de condições de segurança e saúde e o desenvolvimento sustentável.

Figura 7 – Importância relativa dos dez indicadores mais relevantes



Fonte: Elaboração própria (2025).

4.3 Contribuições dos indicadores (SHAP)

Os valores SHAP quantificam a contribuição individual de cada indicador para a predição do modelo, baseando-se na teoria dos jogos cooperativos de Shapley (1953). Cada valor SHAP representa a contribuição marginal de um indicador sobre o IDSC, considerando todas as combinações possíveis de indicadores. Assim, os indicadores com as maiores médias de valores absolutos de SHAP são os que mais contribuem com o resultado do modelo.

O Quadro 2 mostra os dez indicadores com as maiores médias de valores absolutos de SHAP, destacando-se novamente os relacionados com Vida na Água (ODS-14) e água potável e saneamento (ODS-6).

Quadro 2 – Descrição dos indicadores com as maiores médias de valores absolutos de SHAP

(**Continua**)

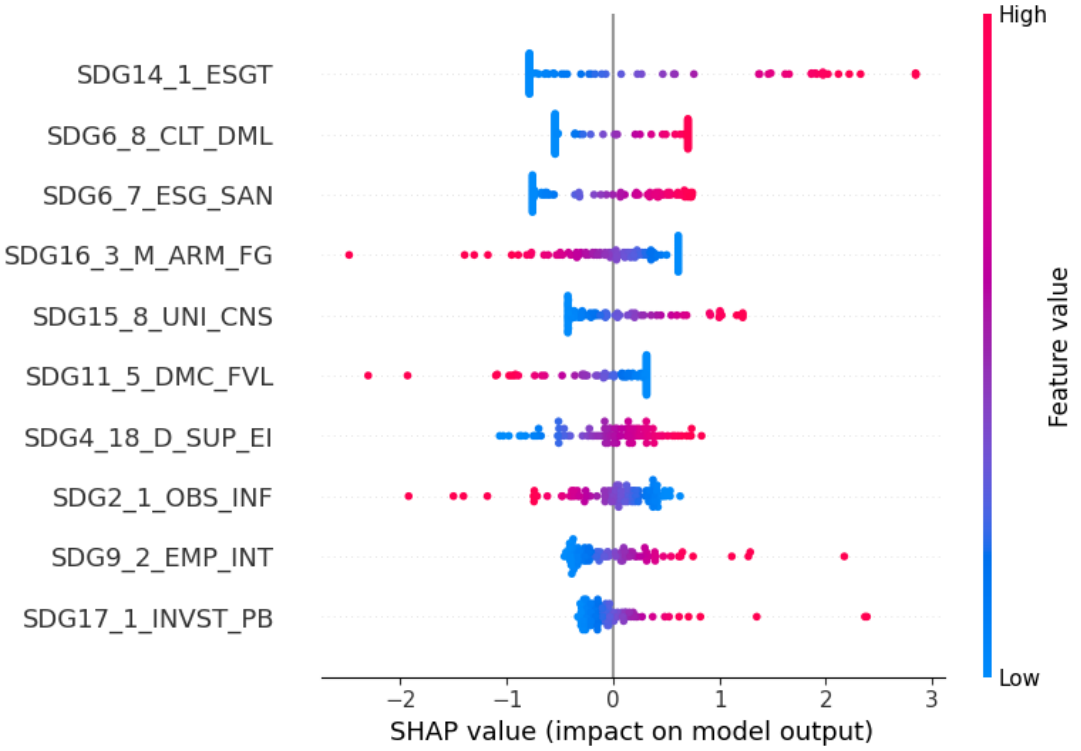
Código	**ODS**	**Indicador**
SDG14_1_ESGT	14 – Vida na Água	Esgoto tratado antes de chegar ao mar, rios e córregos (%)
SDG6_8_CLT_DML	6 – Água Potável e Saneamento	Índice de tratamento de esgoto (%)

Código	ODS	Indicador
SDG6_7_ESG_SAN	6 – Água Potável e Saneamento	População atendida com esgotamento sanitário (%)
SDG16_3_M_ARM_FG	16 – Paz, Justiça e Instituições Eficazes	Mortes por armas de fogo (100 mil habitantes)
SDG15_8_UNI_CNS	15 - Vida Terrestre	Unidades de conservação de proteção integral e uso sustentável (%)
SDG11_5_DMC_FVL	11 – Cidades e Comunidades Sustentáveis	Domicílios em favelas (%)
SDG4_18_D_SUP_EI	4 – Educação de Qualidade	Professores com formação em nível superior - Educação Infantil - rede pública (%)
SDG2_1_OBS_INF	2 – Fome Zero e Agricultura Sustentável	Obesidade infantil (%)
SDG9_2_EMP_INT	9 – Indústria, Inovação e Infraestrutura	Participação dos empregos formais em atividades intensivas em conhecimento e tecnologia (%)
SDG17_1_INVST_PB	17 – Parcerias e Meios de Implementação	Investimento público (R$ *per capita*)

Fonte: Elaboração própria (2025).

A Figura 8 sintetiza a importância e a contribuição dos dez indicadores mais relevantes para a predição do IDSC. No eixo vertical, os indicadores são ordenados pela média dos valores absolutos de SHAP, ou seja, pela sua relevância global na explicação do IDSC. No eixo horizontal, a posição indica a magnitude e o sentido da contribuição: valores SHAP positivos elevam o IDSC predito, enquanto negativos o reduzem. As cores representam o valor original do indicador (vermelho = alto, azul = baixo), permitindo observar como variações nos indicadores afetam a predição. Assim, a Figura 8 revela os indicadores com maior poder explicativo na predição do IDSC do ERJ, segundo o Ridge.

Figura 8 – Os dez indicadores com as maiores médias de valores absolutos de SHAP



Fonte: Elaboração própria (2025).

Observa-se na Figura 8 que o percentual de esgoto tratado antes de chegar ao mar, rios e córregos (SDG14_1_ESGT), o índice de tratamento de esgoto (SDG6_8_CLT_DML) e o percentual da população atendida com esgotamento sanitário (SDG6_7_ESG_SAN) estão positivamente associados ao IDSC, enquanto as mortes por arma de fogo (SDG16_3_M_ARM_FG), o percentual de domicílios em favelas (SDG11_5_DMC_FVL) e o percentual de obesidade infantil (SDG2_1_OBS_INF) estão negativamente associados ao IDSC. Essa análise evidencia que os indicadores com maior poder explicativo na predição do IDSC do ERJ combinam dimensões ambientais, sociais e institucionais, coerentes com a natureza multidimensional da Agenda 2030.

Os indicadores na Figura 7 e na Figura 8 não coincidem totalmente porque cada técnica responde a questões diferentes. Os valores de importância foram obtidos a partir dos coeficientes do Ridge, refletindo a contribuição global de cada indicador segundo métricas lineares tradicionais. Já os valores SHAP consideram a contribuição marginal de cada indicador sobre as predições do Ridge para cada município, indicando a magnitude e o sentido das contribuições.

Assim, o conjunto de indicadores pode divergir porque as importâncias privilegiam a estabilidade, enquanto o SHAP reflete exclusivamente a lógica interna do Ridge. Essa diferença deve ser interpretada de forma complementar: o *ranking* de

importâncias mostra quais indicadores são consistentemente relevantes e o SHAP mostra como esses indicadores contribuem para a predição do IDSC, positiva ou negativamente, no nível individual (municípios).

Portanto, os resultados confirmam a importância de políticas integradas de saneamento, segurança pública, meio-ambiente, inclusão social e saúde para elevar o desempenho sustentável dos municípios do ERJ, corroborando os achados de Crane *et al*. (2021) e Parikh *et al*. (2021) que apontam sinergia entre essas áreas e a relevância do saneamento para o desenvolvimento sustentável. Além disso, a consistência entre as análises de importâncias e SHAP reforça a confiabilidade do modelo de ML selecionado, oferecendo subsídios robustos para a formulação de políticas públicas baseadas em evidências.

A análise realizada possui natureza exclusivamente associativa, não permitindo inferências de causalidade. As relações identificadas expressam a configuração empírica observada no sistema, evidenciando a consistência dos componentes mais relevantes do IDSC. Adicionalmente, a exclusão de parte dos indicadores originais com base em critérios de correlação e VIF implica que o modelo final apresenta uma estrutura distinta da formulação oficial do índice. Nesse sentido, o objetivo não é reproduzir sua metodologia normativa, mas avaliar a robustez e a relevância empírica de seus principais fatores.

5 CONSIDERAÇÕES FINAIS

O presente estudo analisou os fatores associados ao IDSC nos municípios do ERJ, por meio de um *pipeline* robusto que combina imputação de dados ausentes, tratamento de multicolinearidade, estimação de modelos de ML via validação cruzada, comparação e seleção do melhor modelo, identificação de indicadores importantes para a predição do IDSC e a análise da explicabilidade com valores SHAP. A escolha do modelo Ridge, fundamentada em seu maior desempenho médio de R^2 e na estabilidade dos coeficientes, mostrou-se adequada frente aos indicadores que compõem o IDSC.

Os resultados evidenciaram que o percentual de esgoto tratado antes de chegar ao mar, rios e córregos (SDG14_1_ESGT), o índice de tratamento de esgoto (SDG6_8_CLT_DML) e o percentual da população atendida com esgotamento sanitário (SDG6_7_ESG_SAN) foram os três principais indicadores explicativos das predições do

IDSC. Destaca-se também a relevância de indicadores relacionados ao meio-ambiente, inclusão social e saúde para o desempenho sustentável municipal.

A análise de explicabilidade via SHAP permitiu uma compreensão mais aprofundada, ao indicar o sentido e a magnitude das contribuições dos indicadores na predição do IDSC. Observou-se que saneamento e conservação ambiental contribuíram positivamente, enquanto fatores relacionados à violência, desigualdades sociais e más condições de saúde contribuíram negativamente.

Conclui-se, portanto, que a sustentabilidade municipal no ERJ está fortemente associada a dimensões sociais e ambientais interdependentes, o que reforça a necessidade de políticas públicas integradas e multissetoriais. O estudo também demonstra a viabilidade de métodos baseados em ciência de dados para auxiliar gestores públicos na identificação de prioridades e na formulação de estratégias alinhadas à Agenda 2030.

Como limitação desse trabalho, destaca-se a natureza transversal dos dados do IDSC, ao retratar o desempenho sustentável dos municípios em um momento específico, permitindo comparações entre unidades, mas não análises de evolução temporal.

Sugere-se, como trabalhos futuros, a ampliação da análise para outros estados, análise de causalidade e a incorporação de técnicas de séries temporais, a fim de capturar a evolução dinâmica dos indicadores mais explicativos do IDSC ao longo do tempo.

REFERÊNCIAS

AKOGLU, H. User's guide to correlation coefficients. **Turkish Journal of Emergency Medicine**, v. 18, n. 3, p. 91–93, 2018.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. **Biometrika**, v. 76, n. 3, p. 503–514, 1989.

BUUREN, S. van. **Flexible imputation of missing data**. 2. ed. Boca Raton: Chapman and Hall/CRC, 2018.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K. MICE: Multivariate Imputation by Chained Equations in R. **Journal of Statistical Software**, v. 45, n. 3, p. 1–67, 2011.

COSTA, A. P.; FERNÁNDEZ, V. L. Sustainability and development in the municipalities of the State of Paraná: Mapping and analysis using the sustainable city development index of Brazil (IDSC-BR), **CEPAL Review**, v. 2024, n. 143, p. 85–107, 2024.

CRANE, M.; LLOYD, S.; HAINES, A.; DING, D.; HUTCHINSON, E.; BELESOVA, K.; DAVIES, M.; OSRIN, D.; ZIMMERMANN, N.; CAPON, A.; WILKINSON, P.; TURCU, C. Transforming cities for sustainability: a health perspective. **Environment International**, [S.l.], v. 147, a. 106366, 2021.

DUNN, O. J. Multiple comparisons using rank sums. **Technometrics**, v. 6, n. 3, p. 241–252, 1964.

GUIMARÃES, L. Neto. Desigualdades e Políticas Regionais no Brasil: Caminhos e Descaminhos. **Revista Planejamento e Políticas Públicas**. Brasília: IPEA, n. 15, jun. 1997.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009.

HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for nonorthogonal problems. **Technometrics**, v. 12, n. 1, p. 55–67, 1970.

INSTITUTO CIDADES SUSTENTÁVEIS. **Apresentação Índice de Desenvolvimento Sustentável das Cidades - Brasil**. [s.l.], [s.d.-a]. Disponível em: https://idsc.cidadessustentaveis.org.br/introduction/. Acesso em: 08 ago. 2025.

INSTITUTO CIDADES SUSTENTÁVEIS. **Índice de Desenvolvimento Sustentável das Cidades**. [s.l.], [s.d.-b]. Disponível em: https://www.cidadessustentaveis.org.br/paginas/idsc-br. Acesso em: 08 ago. 2025.

INSTITUTO JONES DOS SANTOS NEVES. **Pobreza nos estados brasileiros 2024**. Vitória, ES: IJSN, 2025. Disponível em: https://ijsn.es.gov.br/Media/IJSN/PublicacoesAnexos/S%C3%ADnteses/IJSN_Especial_Pobreza_Estados_Brasileiros_2024_BR.pdf. Acesso em: 29 set. 2025.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with applications in R**. New York: Springer, 2013.

KOLMOGOROV, A. N. Sulla determinazione empirica di una legge di distribuzione. **Giornale dell'Istituto Italiano degli Attuari**, v. 4, p. 83–91, 1933.

KRUSKAL, W. H., WALLIS, W. A. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, v. 47, n. 260, p. 583–621, 1952.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York: Springer, 2013.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. New York: John Wiley & Sons, 1987.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: **Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)**. Long Beach, CA: Curran Associates, p. 4765–4774, 2017.

MARQUARDT, D. W. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. **Technometrics**, v. 12, n. 3, p. 591–612, 1970.

MASSEY, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. **Journal of the American Statistical Association**, v. 46, n. 253, p. 68–78, 1951.

NAÇÕES UNIDAS BRASIL. **Agenda 2030 para o Desenvolvimento Sustentável**. Brasília, 2015. Disponível em: https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustent%C3%A1vel. Acesso em: 08 set. 2025.

PARIKH, P.; DIEP, L.; HOFMANN, P.; TOMEI, J.; CAMPOS, L. C.; TEH, T.; MULUGETTA, Y.; MILLIGAN, B.; LAKHANPAUL, M. Synergies and trade-offs between sanitation and the sustainable development goals. **UCL Open Environment**, London, v. 3, n. 1, 2021.

PEARSON, K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. **Philosophical Transactions of the Royal Society A**, v. 187, p. 253–318, 1896.

PINHEIRO, M. M. S. Políticas públicas baseadas em evidências: um modelo moderado de análise conceitual e avaliação crítica. *In*: KOGA, N. M.; PALOTTI, P. L. D. M.; MELLO, J.; PINHEIRO, M. M. S. (Org.). **Políticas públicas e usos de evidências no Brasil:** conceitos, métodos, contextos e práticas. Brasília: Ipea, 2022. Disponível em: https://portalantigo.ipea.gov.br/agencia/images/stories/PDFs/livros/livros/220412_lv_o_que_informa_miolo_cap01.pdf. Acesso em: 5 set. 2025.

SCIKIT-LEARN. **IterativeImputer**. Scikit-learn: Machine Learning in Python. [s.l.], [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html. Acesso em: 08 set. 2025.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3–4, p. 591–611, 1965.

SHAPLEY, L. S. A value for n-person games. *In*: KUHN, H. W.; TUCKER, A. W. (org.). **Contributions to the theory of games**. Princeton: Princeton University Press, v. 2, p. 307–317, 1953.

SMIRNOV, N. Table for estimating the goodness of fit of empirical distributions. **Annals of Mathematical Statistics**, v. 19, n. 2, p. 279–281, 1948.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B**, v. 58, n. 1, p. 267–288, 1996.

WISSMANN, M. A.; BACKES, G. Índice de Desenvolvimento Sustentável das Cidades: Um Estudo Com Base na Realidade Brasileira. **Revista Científica Acertte**, v. 2, n. 9, p. e2991, 2022.

YANG, K.; WANG, H.; DAI, G.; HU, S.; ZHANG, Y.; XU, J. Determining the repeat number of cross-validation. *In*: **2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)**, Shanghai, China. IEEE, p. 1706–1710, 2011.

O artigo assinado é de inteira responsabilidade dos autores, bem como no que se refere ao uso de imagens.