

Aprendizado de Máquina Explicável para Análise Estrutural do Desenvolvimento Sustentável em Minas Gerais

Leandro Scala da Rocha , Sady Júnior Martins da Costa de Menezes 

Universidade Federal Rural do Rio de Janeiro (UFRRJ) – Três Rios, RJ – Brasil

{scala.leandro, sadymenezes}@ufrrj.br

Abstract. *This study analyzed the Sustainable Development Index of Cities (SDIC) across municipalities in the state of Minas Gerais. The approach combined preprocessing, multicollinearity control, model fitting (Elastic Net, Ridge, Random Forest, and XGBoost) through cross-validation, and performance evaluation. Explainability was conducted using the SHAP (SHapley Additive exPlanations) method. The model with L2 regularization achieved the best performance. The results indicate the relevance of sanitation and public safety indicators, providing empirical evidence to support public policies aligned with the Sustainable Development Goals (SDGs) of the 2030 Agenda.*

Resumo. *Este estudo analisou o Índice de Desenvolvimento Sustentável das Cidades (IDSC) dos municípios do estado de Minas Gerais. A abordagem combinou pré-processamento, controle de multicolinearidade, ajuste de modelos (Elastic Net, Ridge, Random Forest e XGBoost) via validação cruzada e avaliação de desempenho. A explicabilidade foi realizada a partir do método SHAP (SHapley Additive exPlanations). O modelo com regularização L2 apresentou melhor desempenho. Os resultados indicam a relevância de indicadores de saneamento básico e segurança, oferecendo evidências empíricas para apoiar políticas públicas alinhadas aos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030.*

1. Introdução

Diante da emergência climática, a Agenda 2030 estabeleceu metas globais para o desenvolvimento sustentável, por meio dos ODS [Nações Unidas Brasil 2015]. O IDSC consolida indicadores municipais associados aos ODS, permitindo monitorar a sustentabilidade urbana em escala local [Instituto Cidades Sustentáveis s.d.-a]. Nesse contexto, técnicas de Aprendizado de Máquina (AM) e análise de dados vêm sendo utilizadas como suporte à transformação digital da gestão pública orientada por dados [Valle-Cruz *et al.* 2019].

Segundo estimativas do Instituto Brasileiro de Geografia e Estatística (s.d), o estado de Minas Gerais (MG) concentra cerca de 10% da população brasileira distribuída em 853 municípios, o que evidencia a relevância da investigação dos fatores associados ao IDSC em escala regional. Estudos regionais permitem incorporar especificidades territoriais relevantes à formulação de políticas públicas [Guimarães 1997].

Estudos recentes têm utilizado o IDSC principalmente sob perspectiva descritiva. Wissmann e Backes (2022) realizaram comparações entre municípios brasileiros sem utilização de modelagem preditiva. De forma semelhante, Costa e Fernández (2024) analisaram o desempenho dos municípios do Paraná nos ODS.

Este estudo propõe um *pipeline* integrado de ciência de dados para análise estrutural do IDSC dos municípios do MG, englobando modelos de AM e explicabilidade baseada em SHAP.

O objetivo do trabalho é identificar os indicadores mais relevantes associados ao IDSC e analisar sua contribuição para as previsões dos modelos, oferecendo subsídios empíricos para políticas públicas orientadas por dados.

2. Metodologia

A pesquisa adotou abordagem quantitativa aplicada, utilizando dados públicos do IDSC referentes aos anos de 2024 (IDSC-BR_2024) e 2025 (IDSC-BR_2025). O *dataset* original é composto por 100 indicadores associados aos ODS para 5.570 municípios brasileiros [Instituto Cidades Sustentáveis s.d.-b]. Para este estudo, foram selecionados os 853 municípios do MG.

O IDSC foi definido como variável alvo y e os indicadores como variáveis preditoras X . Para reduzir oscilações anuais, utilizou-se a média dos indicadores disponíveis entre os anos 2024 e 2025.

A exclusão dos indicadores foi conduzida de forma sequencial: (i) indicadores com mais de 5% de valores ausentes; (ii) indicadores integrantes de pares com alta correlação linear absoluta ($|r| > 0,9$), mantendo-se, em cada caso, aquele com maior associação ao IDSC [Akoglu, 2018; Pearson, 1896]; e (iii) remoção iterativa de indicadores até que todos os restantes apresentassem *Variance Inflation Factor* (VIF) inferior a 10 [Marquardt, 1970]. Ao término do procedimento, o conjunto foi reduzido a 93 indicadores.

Foram avaliados quatro modelos de AM: *Elastic Net* [Zou e Hastie 2005], Ridge [Hoerl e Kennard 1970], *Random Forest* [Breiman 2001] e XGBoost [Chen e Guestrin 2016]. A seleção de hiperparâmetros foi realizada por *Grid Search* com validação cruzada *k-fold* ($k = 5$), utilizando partição *holdout* de 80% para treinamento e 20% para validação [Burman 1989]. A escolha final considerou simultaneamente o desempenho preditivo e a capacidade de generalização dos modelos.

Em seguida, os modelos foram comparados por meio de validação cruzada repetida utilizando *Repeated K-Fold* ($k = 5$; $B = 100$), totalizando 500 execuções para cada modelo. O desempenho foi avaliado a partir do coeficiente de determinação (R^2), considerando a média e variabilidade das distribuições obtidas.

O tratamento da multicolinearidade foi realizado previamente à validação cruzada com o objetivo de identificar a estrutura global de redundância dos indicadores. Assim, os resultados devem ser interpretados sob perspectiva associativa e exploratória, podendo existir viés otimista de desempenho decorrente dessa etapa metodológica.

A identificação dos indicadores mais relevantes foi realizada com base no método SHAP [Lundberg e Lee 2017], permitindo analisar magnitude e sentido da contribuição individual dos indicadores nas previsões do IDSC.

3. Resultados e Discussão

A Figura 1 apresenta a distribuição do R^2 em 500 execuções da validação cruzada repetida para os modelos avaliados. Observa-se desempenho superior do *Elastic Net*, Ridge e

XGBoost em relação ao *Random Forest*, por apresentarem maiores medianas e menores dispersões.

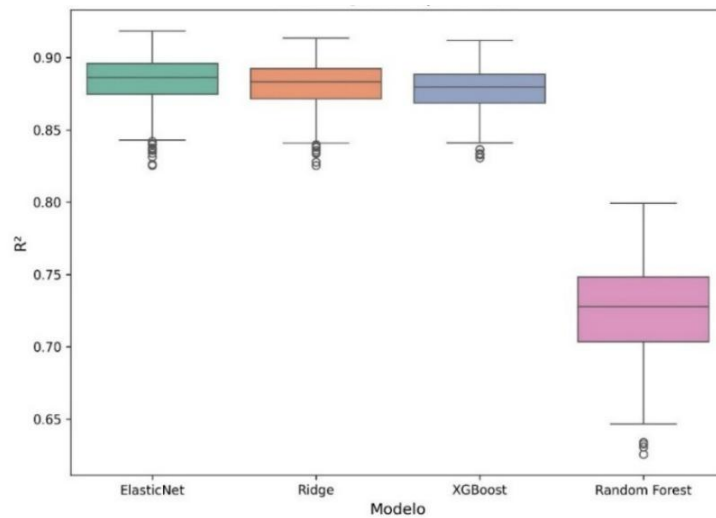


Figura 1. Variabilidade do R^2 em 500 simulações.

O teste de Shapiro–Wilk [Shapiro e Wilk 1965] indicou violação da hipótese de normalidade da distribuição do R^2 ($p < 0,05$), motivando a utilização do teste não paramétrico de Kruskal–Wallis [Kruskal e Wallis 1952], seguido de comparações *post hoc* de Dunn com correção de Bonferroni [Dunn 1964]. Entretanto, devido à dependência entre as observações geradas pela validação cruzada repetida, tais análises devem ser interpretadas de forma exploratória.

A Tabela 1 apresenta as métricas de desempenho dos modelos. O *Elastic Net* obteve o melhor desempenho médio. A razão L1 ótima do *Elastic Net* foi igual a zero, tornando-o equivalente à regressão Ridge em termos de regularização L2.

Tabela 1. Métricas de desempenho dos modelos.

Modelo	R^2 (média)	R^2 (desvio padrão)
<i>Elastic Net</i>	0,8838	0,0164
Ridge	0,8805	0,0160
XGBoost	0,8784	0,0144
<i>Random Forest</i>	0,7256	0,0317

O Quadro 1 mostra os dez indicadores com as maiores médias de valores absolutos de SHAP em ordem decrescente, destacando-se: Esgoto tratado antes de chegar ao mar, rios e córregos (SDG14_1_ESGT), Taxa de homicídio (SDG16_4_TX_HMC) e Índice de tratamento de esgoto (SDG6_8_CLT_DML), evidenciando a relevância do saneamento básico e segurança pública.

Quadro 1. Descrição dos indicadores com as maiores médias de |SHAP|.

Código	ODS	Indicador
SDG14_1_ESGT	14 – Vida na Água	Esgoto tratado antes de chegar ao mar, rios e córregos (%)
SDG16_4_TX_HMC	16 – Paz, Justiça e Instituições Eficazes	Taxa de homicídio (100 mil habitantes)

SDG6_8_CLT_DML	6 – Água Potável e Saneamento	Índice de tratamento de esgoto (%)
SDG11_5_EQP_ESP	11 – Cidades e Comunidades Sustentáveis	Equipamentos esportivos municipais (100 mil habitantes)
SDG17_1_INVST_PB	17 – Parcerias e Meios de Implementação	Investimento público (R\$ <i>per capita</i>)
SDG8_3_T_DSC	8 – Trabalho Decente e Crescimento Econômico	Desemprego (taxa)
SDG7_3_VNLENER	7 – Energia Limpa e Acessível	Vulnerabilidade Energética
SDG15_8_UNI_CNS	15 – Vida Terrestre	Unidades de conservação de proteção integral e uso sustentável (%)
SDG7_2_ENRG	7 – Energia Limpa e Acessível	Domicílios com acesso à energia elétrica (%)
SDG13_PDAR	13 – Ação Contra a Mudança Global do Clima	Proporção de domicílios em áreas de risco

A Figura 2 apresenta os valores SHAP dos dez indicadores mais relevantes para o *Elastic Net*. Observa-se a associação positiva entre os indicadores de saneamento e os valores preditos do IDSC e o inverso com a taxa de homicídio.

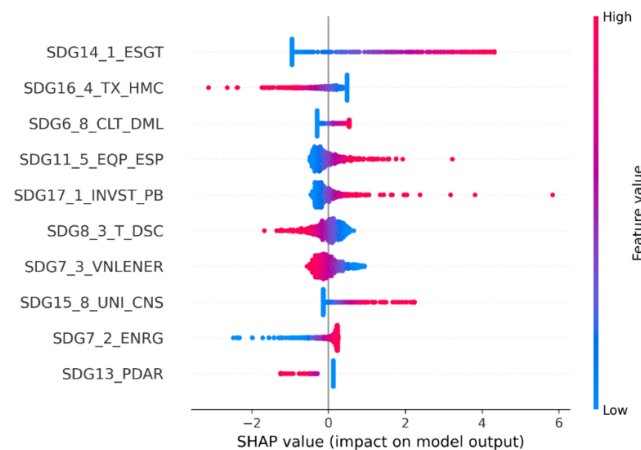


Figura 2. Top 10 valores SHAP (*Elastic Net*).

Os resultados reforçam a relevância de políticas públicas integradas nas áreas de saneamento básico, segurança e infraestrutura, em consonância com Crane *et al.* (2021) e Parikh *et al.* (2021), que destacam sinergias entre esses setores.

A análise possui caráter estritamente associativo e não causal. Além disso, como a redução de indicadores foi realizada antes da validação cruzada, os resultados refletem a capacidade preditiva dos modelos em um conjunto previamente estruturado de indicadores, e não uma validação completamente fora da amostra.

4. Conclusão

Este estudo analisou os principais fatores associados ao IDSC dos municípios do MG, por meio de um *pipeline* integrado de ciência de dados. A abordagem combinou tratamento de dados, controle de multicolinearidade, ajuste de modelos, validação cruzada repetida e explicabilidade baseada em SHAP.

Os resultados indicaram desempenho superior dos modelos com regularização L2, destacando a relevância de indicadores relacionados ao saneamento básico, segurança pública, preservação ambiental e infraestrutura urbana. A análise de explicabilidade baseada em SHAP permitiu interpretar tanto a importância global quanto o impacto local dos indicadores sobre as previsões do IDSC.

O estudo demonstra o potencial da integração entre AM e análise de indicadores territoriais como suporte à gestão pública orientada por dados e aos ODS da Agenda 2030.

Como limitações, destacam-se o caráter transversal dos dados e a realização da redução de indicadores antes da validação cruzada, o que pode introduzir viés otimista de desempenho. Pesquisas futuras podem ampliar a abordagem para outras regiões, incorporar técnicas mais recentes de AM, validação externa e métodos de inferência causal.

Referências

- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Breiman, L. (2001). *Random forests*. *Machine Learning*, 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- 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. <https://doi.org/10.2307/2336116>
- Costa, A. P., & Fernández, V. L. (2024). 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*, 2024(143):85–107. <https://doi.org/10.18356/16840348-2024-143-5>
- Chen, T., & 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. Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Crane, M., Lloyd, S., Haines, A., Ding, D., Hutchinson, E., Belesova, K., Davies, M., Osrin, D., Zimmermann, N., Capon, A., Wilkinson, P., & Turcu, C. (2021). Transforming cities for sustainability: a health perspective. *Environment International*, 147, 106366. <https://doi.org/10.1016/j.envint.2020.106366>
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252. <https://doi.org/10.1080/00401706.1964.10490181>
- Guimarães, L. Neto (1997). Desigualdades e políticas regionais no Brasil: caminhos e descaminhos. *Revista Planejamento e Políticas Públicas*, 15. <https://repositorio.ipea.gov.br/entities/publication/fea3d425-09b7-4d6b-be2d-879ac5ea0e2>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>

- Instituto Brasileiro de Geografia e Estatística. (s.d.). *Cidades e Estados do Brasil*. <https://cidades.ibge.gov.br/>. Acesso em: 28 mar. 2026.
- Instituto Cidades Sustentáveis. (s.d.-a). *Apresentação índice de desenvolvimento sustentável das cidades – brasil*. Índice de Desenvolvimento Sustentável das Cidades – Brasil. <https://idsc.cidadessustentaveis.org.br/introduction>. Acesso em: 15 set. 2025.
- Instituto Cidades Sustentáveis. (s.d.-b). *Índice de desenvolvimento sustentável das cidades*. <https://www.cidadessustentaveis.org.br/paginas/idsc-br>. Acesso em: 15 set. 2025.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pages 4765–4774. Curran Associates. https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612. <https://doi.org/10.1080/00401706.1970.10488699>
- Nações Unidas Brasil. (2015). *Agenda 2030 para o desenvolvimento sustentável*. <https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustent%C3%A1vel>.
- Parikh, P., Diep, L., Hofmann, P., Tomei, J., Campos, L. C., Teh, T., Mulugetta, Y., Milligan, B., & Lakhanpaul, M. (2021). Synergies and trade-offs between sanitation and the sustainable development goals. *UCL Open Environment*, 2, e016. <https://doi.org/10.14324/111.444/ucloe.000016>
- Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, (187):253–318. <https://doi.org/10.1098/rsta.1896.0007>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Valle-Cruz, D., Ruvalcaba-Gómez, E. A., Sandoval-Almazán, R., & Criado, J. I. (2019). A review of artificial intelligence in government and its potential from a public policy perspective. *Proceedings of the 20th Annual International Conference on Digital Government Research*, 91–99. <https://doi.org/10.1145/3325112.3325242>
- Wissmann, M. A., & Backes, G. (2022). Índice de desenvolvimento sustentável das cidades: um estudo com base na realidade brasileira. *Revista Científica Acertte*, 2(9), e2991. <https://doi.org/10.47820/acertte.v2i9.91>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, London, 67(2):301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>