Cleaner Production Letters 8 (2025) 100088

Contents lists available at ScienceDirect

Cleaner Production Letters

journal homepage: www.journals.elsevier.com/cleaner-production-letters

Sustainable development goal 6 monitoring through statistical machine learning – Random Forest method

Murilo de Carvalho Marques [a], Abdoulaye Aboubacari Mohamed [b], Paulo Feitosa [c,*]

[a] University of São Paulo, Av. Pádua Dias, 11, Piracicaba, SP, 13418-900, Brazil
[b] Department of Rural Economics, Universidade Federal de Viçosa, Avenida Purdue, s/nº, Campus Universitário, Edifício Edson Potsch Magalhães, 36570-900, Viçosa, Minas Gerais, Brazil
[c] University of São Paulo, Av. Prof. Lúcio Martins Rodrigues, 443, Butantã, São Paulo, SP, Brazil

ARTICLE INFO

Keywords:
Machine learning
SDG 6
Ecosystem monitoring
Random forest
Satellite imagery
Protected areas

ABSTRACT

Global reports from the United Nations project significant deficits in achieving water and sanitation targets by 2030, emphasizing the need for advanced methodologies in ecosystem monitoring. This study examines the integration of the Random Forest machine learning algorithm with freely available satellite imagery and open-source tools to monitor Permanent Protected Areas (PPAs) in the Distrito Federal, Brazil, contributing to Sustainable Development Goal (SDG) 6, which prioritizes clean water and sanitation. The research adopts a methodological approach that classifies land use changes within PPAs, with a focus on riparian zones along riverbanks, utilizing high-resolution Sentinel-2 satellite data processed through the Google Earth Engine platform. The findings indicate a 6% increase in native vegetation within PPAs from 2015 to 2022, highlighting the utility of machine learning technologies in environmental monitoring. The Random Forest algorithm demonstrated robust performance, with classification accuracy rates ranging from 83% to 88% and Kappa coefficients between 0.73 and 0.84. These results underscore the method's ability to enhance data granularity and reliability, supporting informed decision-making in ecosystem management. This research contributes to advancements in environmental monitoring methodologies and aligns with international efforts to achieve SDG targets. Further studies should investigate the incorporation of additional machine learning models to improve monitoring accuracy and support sustainable development initiatives.

1. Introduction

In the context of global sustainability efforts, the United Nations (UN) member states' adoption of the Sustainable Development Goals (SDGs) in 2015 represented a significant pledge to eliminate poverty, protect the environment, and achieve sustainable development by the year 2030. SDG 6 plays a pivotal role in this agenda, underlining the essential nature of clean water and sanitation for enhancing human well-being, driving economic and social progress, and preserving vital ecosystems (Bebbington and Unerman, 2018; Vazquez-Brust et al., 2020). Addressing SDG 6 is fundamental to the broader objectives of sustainability, given the intricate challenges of water management and the need to protect ecosystems critical to water security. These challenges, heightened by the diverse nature of water-related ecosystems in various locales, demand customised strategies for effective management and preservation (Chen and Liu, 2019; Madrazo-Ortega and Molinos-Senante, 2023; Miao et al., 2023; Mustafa et al., 2022).

The United Nations Sustainable Development Goals Report (2022) paints a stark picture of the current trajectory towards achieving SDG 6. At the present rate of progress, an estimated 1.6 billion individuals will remain without access to safe drinking water, 2.8 billion will be without safe sanitation facilities, and 1.9 billion will lack basic hand hygiene by 2030. To align with the objectives of SDG 6, the rate of improvement must be quadrupled (Miao et al., 2023). These alarming forecasts underscore the urgent need for enhanced methodologies and innovative approaches in monitoring and improving water management and sanitation practices (Fuente et al., 2020; Nkiaka et al., 2021).

To expedite the implementation of SDG 6, the United Nations initiated the "SDG 6 Global Acceleration Framework", encompassing five integral domains: finance, data and information, capacity development, innovation, and governance. A critical examination of the data and information domain reveals a significant challenge in the collection of

* Corresponding author. University of São Paulo, Av. Prof. Lúcio Martins Rodrigues, 443, Butantã, São Paulo, SP, 05508-0203, Brazil.
E-mail addresses: murilo.c.marques@outlook.com (M. Carvalho Marques), abdeltoure2229@gmail.com (A.A. Mohamed), pfeitosa@usp.br (P. Feitosa).

https://doi.org/10.1016/j.clpl.2024.100088
Received 30 April 2024; Received in revised form 3 December 2024; Accepted 13 December 2024
Available online 14 December 2024

essential monitoring data (Arora and Mishra, 2022; Miao et al., 2023; Pigola et al., 2021; Hino et al., 2018; Jiang et al., 2024). According to the SDG 6 Progress Summary Report (2021), although most of the 193 UN member states have amassed two-thirds of the required SDG 6 monitoring data, 24 countries have yet to collect half of the necessary information. The scarcity of data is particularly pronounced in smaller regions below the national level, underscoring the urgency of enhancing data acquisition methods to monitor and support the achievement of SDG 6 targets effectively.

The overarching challenge in the domain of sustainable water management and ecosystem preservation under SDG 6 lies in the inadequate monitoring capabilities that hinder the effective conservation of water-related ecosystems, especially in smaller, localized regions (Denu et al., 2023). Despite the global commitment to sustainable development, the lack of comprehensive, high-quality data poses a significant barrier to assessing progress accurately and implementing targeted interventions (Guo et al., 2023; Hofmann, 2021; Nilashi et al., 2023; Sheffield et al., 2018). This gap in data collection and monitoring capabilities is particularly acute in the context of the protected areas, where the intricate dynamics of water ecosystems require nuanced understanding and management. The inability to gather and analyse detailed environmental data at such a granular level compromises the global efforts towards achieving the water-related targets of SDG 6, underscoring the need for innovative solutions that can bridge these gaps and facilitate informed decision-making for the preservation of vital ecosystems.

The existing literature synthesizes diverse methodologies and outcomes in environmental sustainability, notably in water treatment and wastewater management, underlining their significant contributions towards achieving clean water (SDG 6) and promoting responsible consumption (SDG 12) (K. Zhang et al., 2023; W. Zhang et al., 2023; Zhu et al., 2023). Furthermore, the optimization of urban-agricultural-ecological spaces (Wang et al., 2022) and the innovative conversion of wet waste into energy resources (Zhu et al., 2023) illustrate the pivotal role of machine learning (ML) in advancing sustainable cities (SDG 11) and clean energy (SDG 7). Additionally, ML's application in monitoring poverty (Alsharkawi et al., 2021) and predicting human development indices (Ramos et al., 2018) underscores its potential impact on eradicating poverty (SDG 1) and enhancing quality education (SDG 4), thereby contributing to a broader understanding and implementation of SDGs.

Despite the growing interest among scholars in studying the SDGs, there remains a notable deficiency in adequate methods for assessing these goals effectively (Denu et al., 2023; Yao and Li, 2023). This study seeks to bridge this gap by leveraging the advancements in interpretable ML, which offers enhanced transparency in the analysis. Specifically, the Random Forest (RF) method is employed to quantify the contribution of each variable towards the preservation of water-related ecosystems, addressing the shortcomings related to the lack of interpretability in traditional models (Lin et al., 2023). The interpretability of RF models offers a substantial advantage, facilitating a comprehensive understanding of variable importance in land cover classification. This capability supports enhanced model transparency, enables informed decision-making, and ensures reliability by addressing uncertainties associated with environmental data analyses (Meyer et al., 2019; Nilashi et al., 2023; Fisher et al., 2024). Through this approach, the research aims to provide a more nuanced understanding and evaluation of SDG 6.6, thereby contributing to the broader field of sustainable development research with a method that combines robustness with interpretability.

Employing the Random Forest method to monitor the preservation of the most relevant water-related ecosystem in the *Distrito Federal* (Brazil): the riparian areas of its rivers which are strips of moisture-loving vegetation growing along the edge of the water bodies. The Brazilian Forest Code, Law nº. 12,651 of May 25, 2012, mandates the protection of riparian areas. These protected areas are referred to as Permanent Protected Areas (PPAs). This study classified images from the Sentinel-2 satellite, observing an approximate 6% increase in native vegetation within the PPAs of the *Distrito Federal* between 2015 and 2022. The accuracy of these observations ranged between 83% and 88%, with a Kappa coefficient between 0.73 and 0.84, both closely aligning with values reported in the literature. This research utilised freely available datasets, open-source software, and a cloud-based platform: Google Earth Engine (GEE), to conduct its analysis. Through this methodology, the study contributes to the field by providing a replicable model for monitoring ecosystem preservation efforts, leveraging advanced statistical ML techniques within an accessible technological framework.

2. Literature review

2.1. Protect and restore water-related ecosystems (SDG 6.6)

The Sustainable Development Goals (SDGs) were established in 2015 by the United Nations (UN) as a global endeavour to eradicate poverty, protect the environment, and ensure sustainable development by the year 2030. Among these, SDG 6, known for its focus on "Clean Water and Sanitation", goes beyond its primary objective to include targets aimed at evaluating the social, environmental, and economic significance of water resources (Diep et al., 2021; Taka et al., 2021). This goal is detailed through eight targets and 11 indicators, providing a comprehensive framework for assessing progress (Madrazo-Ortega and Molinos-Senante, 2023; Requejo-Castro et al., 2020). However, Basu and Dasgupta (2021) have identified a disconnection between scientific research and the practical application of SDG 6, highlighting a growing divide between technological advancements, academic discussions, and the actual implementation by policy practitioners. This gap underscores the necessity for clearer integration of research findings with policy-making processes to enhance the effectiveness of SDG 6 initiatives.

Evaluating progress and efficiency in local governance is crucial for the achievement of SDG 6, particularly in ensuring access to clean water and sanitation. Martínez-Córdoba et al. (2020) highlight the significant role of Spanish local governments in this endeavour, identifying key determinants such as water-related rates and the private management of water services that enhance governance efficiency. The research conducted by Robins et al. (2017) offers a pertinent perspective on strengthening water governance in complex, multi-tiered arrangements, focusing on the United Kingdom's collective water policies and approaches. This analysis is foundational for understanding the importance of governance in water and sanitation projects. Furthermore, Ba et al. (2022) assess the regulatory challenges faced by the Niger Basin Authority, particularly the lack of legal instruments for effluent discharge, which directly impacts water quality under SDG 6.3. This evaluation underscores the critical need for robust legal frameworks to support the objectives of SDG 6, indicating that effective governance and comprehensive legal regulations are essential for sustainable water management.

The significance of public awareness and community participation in achieving SDG 6 cannot be overstated, as evidenced by the research conducted by Mustafa et al. (2022). Their findings underscore the impact of public awareness, alongside other factors, on advancing SDG 6 in a developing country context, advocating for enhanced strategies to boost public engagement. This underscores the pivotal role of public awareness in progressing towards SDG 6 (Mustafa et al., 2022). Furthermore, the evaluation by Partzsch et al. (2021) of coffee certification standards and their contribution to watershed sustainability light on the environmental emphasis of such programs. However, it also points out their shortcomings in addressing the broader needs of water and sanitation, thereby exploring the influence of industry standards in promoting SDG 6 (Partzsch et al., 2021).

In the domain of scientific research, collaboration, and the exploration of emerging concepts are crucial for advancing SDG 6. Basu and

Dasgupta (2021) conducted a bibliometric analysis that unveils research trends and identifies gaps in the field of water sustainability, calling for increased collaboration and investigation into new concepts. This analysis illuminates the scientific community's significant contributions towards SDG 6. Moreover, the employment of multisource data by Miao et al. (2023) in Lincang City and the holistic approach adopted by Ortega and Senante (2023) in Chile for quantifying SDG 6 indicators exemplify innovative methodologies for assessing progress. These studies underscore the importance of comprehensive evaluations and the utilisation of diverse data sources in effectively addressing SDG 6.

Expanding the dialogue, the exploration into private sector collaboration for sustainable development reveals the complexities of collaborative governance mechanisms. Through a systematic literature review, the study conceptualises the core dimensions of collaborative governance—hierarchy, formalization, centralization—and the factors affecting the impact of governance choices on sustainable development outcomes. The results indicate that various collaboration types act as governance mechanisms to advance Sustainable Development Goals, suggesting the need for combining different governance arrangements. The effectiveness of collaborative efforts is determined not only by governance dimensions but also by the specific SDG targeted and the nature of the partners involved (Vazquez-Brust et al., 2020). This underscores the intricacies and detailed planning required to formulate efficient collaborative strategies for sustainable development.

Addressing the governance challenges and indicator gaps is crucial for the effective monitoring and implementation of SDG 6 (Bhaduri et al., 2016; Fu et al., 2019). Herrera (2019) and Guppy et al. (2019) provide insightful analyses into the governance difficulties and the inadequacies within the SDG 6 indicator framework, offering potential solutions to mitigate these issues. Their work underscores the complexities involved in tracking and achieving SDG 6. Furthermore, scholars have put forth recommendations for African countries to intensively mobilise resources to ensure universal Water, Sanitation and Hygiene (WASH) services by 2030, presenting a critical viewpoint on the international efforts needed to fulfil SDG 6 (Fuente et al., 2020; Hofmann, 2021; Nhamo et al., 2019). This comprehensive approach highlights the multifaceted strategies required to address the challenges of clean water and sanitation globally (Pereira and Marques, 2021).

The literature review underscores the necessity of a multifaceted approach to address the challenges of ensuring clean water and sanitation, as outlined in SDG 6. Key findings highlight the critical role of integrated strategies, encompassing robust governance, comprehensive legal frameworks, heightened public awareness, and rigorous scientific research. These elements are indispensable for the successful achievement of SDG 6, demonstrating the complexity and interconnectivity of efforts required to secure water and sanitation for all.

2.2. Applications with machine learning

The SDGs adopted by UN members in 2015, represent a global commitment to end poverty, protect the planet, and ensure sustainable development by 2030 (UN, 2015). Among these, SDG 6 focuses on ensuring the availability and sustainable management of water and sanitation for all, encompassing eight targets and 11 indicators. However, progress towards these goals varies significantly by income level and location, highlighting the need for a nuanced understanding of regional water resource management and policy design for water-related ecological conservation. This variability underscores the importance of contextualizing sustainable development within local realities, as each community exists within a unique socio-economic and geographical landscape (Madrazo-Ortega and Molinos-Senante, 2023; Miao et al., 2023). Moreover, Basu and Dasgupta (2021) reveal that the linkage between scientific research and SDG 6 is often ambiguous, widening the gap between technological advancements, academic discourse, and practical policy implementation.

ML has been recognised as a crucial technology for promoting sustainable development, offering innovative solutions across a range of SDGs. The European Commission has initiated a forward-thinking plan aimed at fostering the use of green digital technologies to benefit the environment through systemic changes known as 'twin transitions' (European Commission, 2021). This approach highlights the substantial potential of ML in facilitating environmental sustainability. For instance, scholars have highlighted ML's contributions towards improving water treatment processes and enhancing construction and demolition waste management, directly supporting goals like clean water (SDG 6) and responsible consumption (SDG 12) (K. Zhang et al., 2023; W. Zhang et al., 2023). Additionally, ML has been applied to optimise urban-agricultural-ecological spaces and convert wet waste into energy resources, showcasing ML's role in promoting sustainable cities (SDG 11) and clean energy (SDG 7) (Wang et al., 2022; Zhu et al., 2023).

Expanding upon the contributions of ML to environmental sustainability, its applications also significantly impact objectives like eradicating poverty (SDG 1) and enhancing quality education (SDG 4). Research has employed ML to monitor poverty levels and forecast human development indices, demonstrating its capacity to tackle critical social challenges (Alsharkawi et al., 2021; Ramos et al., 2018). Additionally, studies examining generational shifts towards sustainability underscore ML's role in identifying societal trends in sustainable lifestyles and job-seeking behaviours, marking a notable pivot towards sustainability across generations (Yamane and Kaneko, 2021). In the context of sustainable urban planning, ML has been shown to support the development of Smart Cities, improving urban management to foster more inclusive, safe, resilient, and sustainable environments. This approach is particularly relevant in addressing contemporary challenges, such as the COVID pandemic, illustrating ML's broad applicability in furthering the SDGs (Heras et al., 2020; Jain et al., 2023).

ML applications permeate the domains of policy formulation, decision-making processes, and the evaluation of SDGs performance. Research demonstrates ML's ability to provide critical insights into environmental sustainability and support the aggregation of evidence for informed policy decisions (Porciello et al., 2020; Yao and Li, 2023). Furthermore, studies highlight the significant impact of ML, particularly Automated Machine Learning (AutoML), on the evaluation and prediction of SDG achievements. This underscores its predictive power as a crucial tool in strategic planning, essential for the successful fulfilment of SDGs (Singpai and Wu, 2020). In hydrological assessments, machine learning models, including Boosted Regression Tree (BRT), Classification and Regression Tree (CART), and RF, have demonstrated considerable applicability in mapping groundwater potential (Al-Obeidat et al., 2015; Naghibi et al., 2016; Otukei and Blaschke, 2010; Phan et al., 2020). Furthermore, the transformative potential of Deep Learning in remote sensing applications has been highlighted, particularly its ability to extract multiscale and multilevel features for precise environmental mapping and prediction, thereby advancing SDG monitoring frameworks (Yuan et al., 2020).

Following the exploration of ML pivotal role in policy and SDG performance assessment, it's essential to delve into the comparative effectiveness of advanced ML techniques in sustainable development. Techniques such as Artificial Neural Networks (ANN), LightGBM, Automated Machine Learning (AutoML), and the Shapley additive explanation (SHAP) technique have been employed with varying degrees of success across different sustainability studies (Yao and Li, 2023). These methods have demonstrated innovative applications, from environmental sustainability to economic and social development, underscoring ML's versatility. The literature review consolidates the critical role of ML in propelling sustainable development across diverse domains, highlighting the necessity for ongoing research and innovation in ML technologies to maximize their contribution to the SDGs. This reiteration emphasizes the transformative potential of ML in achieving global sustainability objectives, advocating for continued advancements and application of ML techniques in addressing the complex challenges

of the SDGs.

3. Material and methods

3.1. Research context

The selected research area is the *Distrito Federal*, one of the 27 states of Brazil and the smallest in terms of area. Situated in the Central-West region, it is renowned for housing the nation's capital, Brasília, as shown in Fig. 1 below. The DF spans an area of 5760.784 square kilometers (km^2), nestled between the states of *Goias* and *Minas Gerais*. The focus of the land use classification was narrowed down to the PPAs of all water bodies within the DF's hydrography, excluding reservoirs and lakes, culminating in a total study area of 208.3 km^2.

In accordance with the Brazilian Forest Code, established by Law No. 12,651 of May 25, 2012, marginal strips ranging from 30 to 500 m are set based on the width of the rivers, designating them as PPAs. This law defines PPAs as protected areas, whether covered by native vegetation or not, with the environmental function of preserving water resources, landscapes, geological stability, and biodiversity, as well as protecting the soil, fauna, and flora, thereby ensuring the well-being of the population.

Rivers less than 10 m in width are required to have a 30-m marginal strip; watercourses ranging from 10 to 50 m in width should have a 50-m strip, and so on. Given that the aim of this study is to evaluate the use of statistical ML in monitoring SDG 6.6, all PPAs were considered with a 30-m marginal strip, aligning with the objectives of this work.

3.2. Data collection

In the methodology's data collection section, the study primarily utilised the 2015 Hydrography data from the *Distrito Federal* Environmental Information System (SISDIA) - Hydrography CRH (2015), which serves as the main public and free environmental database for the *Distrito Federal* (SISDIA, 2023). Initially, the hydrography data from the National Water and Sanitation Agency (ANA) and its Hydrographic Ottocodificada Base 2013 were considered due to its comprehensive geographic information layers, including drainage sections, hydrographic contribution areas, and watercourses. However, ANA's hydrographic base, developed from hydrographic contribution areas derived from the Shuttle Radar Topography Mission (SRTM) project's digital elevation models (DEM) with spatial resolutions of 90 and 30 m depending on the region, resulted in a line-type shapefile that inaccurately represented river boundaries, occasionally extending through urban or agricultural areas outside the water bodies and PPAs. In contrast, SISDIA's database utilises DEMs with higher precision, specifically developed for the *Distrito Federal*.

For delineating the PPAs, the multiplatform open-source geographic information system software QGIS 3.32 was employed, using the BUFFER tool on the line-type Hydrography feature class shapefile obtained from SISDIA, as shown in Fig. 2 above. This tool generates an area around a point, line, or polygon shapefile, allowing for the specification of the width of the marginal strip to be created, in this case, set to 30 m. Additionally, the 2022 states metadata for Brazil, available on the Brazilian institute of Geography and Statistics (IBGE, 2023) portal, were used to define Brazil's, especially the boundaries of the *Distrito Federal*.

Regarding satellite images, this study follows previous research (Radoux et al., 2016) and opts for the Sentinel-2 satellite images over the commonly used Landsat images for several reasons: (i) The superior quality of images due to the smaller pixel size of Sentinel-2, which varies between 10 and 60 m, providing images with a resolution of 10 m for the *Distrito Federal*, compared to Landsat's 30-m pixel. (ii) The study area's specificity, as the analysis focuses on land use within PPAs with 30 m on each side of water bodies, where a higher number of pixels allows for more detailed interpretation and classification accuracy. Using Landsat would result in only 1 pixel per 30 m of river, whereas Sentinel provides 9 pixels, i.e., 9 data/information points for the same area, instead of just one. (iii) The period of the satellite's launch, which coincides with the year the SDGs were established, in 2015, enabling a comparison of the preservation of PPAs at the start of the program between 2015 and 2016, with the most recent data between 2022 and 2023, depending on data availability and quality. This approach offers an overview of the preservation of water-related ecosystems in *Distrito Federal* since the inception of the SDGs, aligning with the objectives of this work.

3.3. Random Forest classification: model training and validation

In this study, the Random Forest algorithm was deployed to classify the land use of the PPAs in *Distrito Federal* through the Google Earth Engine (GEE) platform. The GEE classifier package facilitates supervised classification using traditional ML algorithms, including Classification and Regression Trees (CART), Random Forest, Naive Bayes, and Support Vector Machine (SVM). The "Guides" section on the Google Developers site provides examples and instructions for these algorithms. GEE is widely used in environmental analyses at regional and global scales, with a significant increase in studies since 2017 (Tamiminia et al., 2020).

RF was chosen for this analysis because it is a nonparametric method that does not depend on prior knowledge of the ecological factors or attributes related to the prediction or classification outcomes (Menze et al., 2009). Additionally, like other Classification and Regression Tree (CART) methods, RF offers computational simplicity and can handle large datasets effectively. RF and other CART methods achieved an overall accuracy of 86% in classifying Landsat satellite images

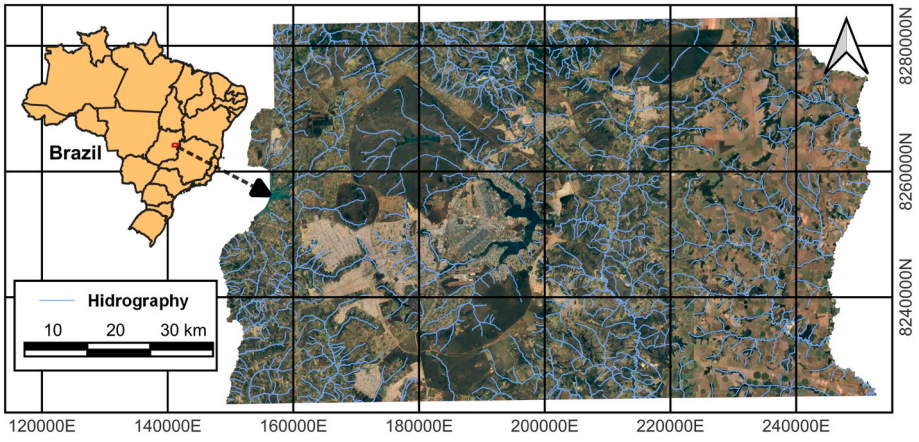

Fig. 1. Map of the study area and its location within Brazil.

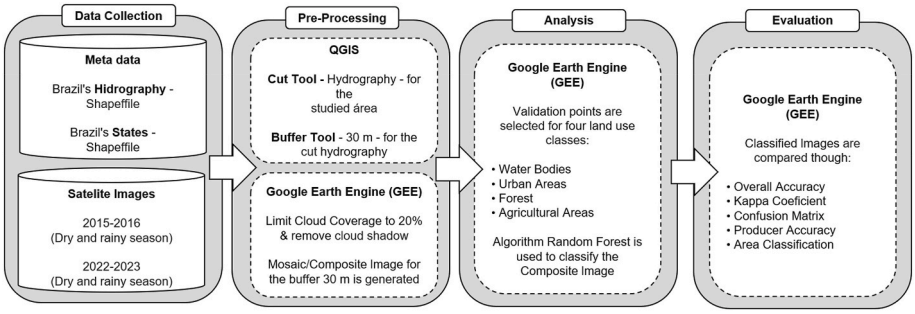

Fig. 2. Workflow diagram.

(Al-Obeidat et al., 2015). Similar findings were reported by Otukei and Blaschke (2010) and Phan et al. (2020), with overall accuracies surpassing 85% and 84% for satellite image classification, respectively.The algorithm employed in this research was based on models available on the Google Developers platform, with modifications such as changing the satellite used, selecting bands, determining the number of classes for classification, and specifying the polygon-type feature class shapefile for supervised classification, specifically targeting the PPAs of water bodies in the DF.

According to Congalton and Green (2019), fifty randomly allocated validation points are deemed sufficient for each land use class. However, their studies typically utilised at least two hundred points per satellite image (mosaic). For model training and validation, 60 points (markers in GEE) were used for each land use class, comprising four categories for the model's classification: (i) LC_Water: Water bodies. (ii) LC_Degradation: Urban areas. (iii) LC_Woody: Forest/native vegetation. (iv) LC_Agriculture: Agricultural areas, as shown in Fig. 3 below:

To guide the placement of validation points (response or output variables) for each class, Sentinel-2 satellite images from 2015 to 2016 and 2022–2023 were compared with Airbus CNES 2023 images available on the GEE platform, positioning them only in areas where land use remained unchanged. For the LC_Woody class, sections of the Mineral Water National Park, a preserved area with only native vegetation, were chosen; for the LC_Water class, wider rivers whose area tends not to change were selected; for the LC_Degradation class, consolidated urban areas were chosen; and for LC_Agriculture, predominantly agricultural locations in the DF were selected, including different crops at various stages of production, as shown in Figs. 4 and 5. After training, the model was applied to the entire sample (a 30-m BUFFER along the margins of water bodies in the DF), yielding the results presented in the next section. This structured approach ensures a comprehensive and coherent methodology for the classification and validation process, adhering to established practices in the field of remote sensing and machine learning.

4. Findings and discussion

To achieve the objective of monitoring SDG 6.6 through statistical ML, focusing on the preservation of PPAs in *Distrito Federal*, images from two periods were sought: shortly after the creation of the SDGs, between 2015 and 2016, and a more recent period, between 2022 and 2023. Utilizing the GEE, it was possible to search for Sentinel-2 satellite images by specifying only the period of interest. The Sentinel satellite was chosen primarily for its high-resolution capabilities. In addition to specifying the search date, filters were applied to limit the maximum cloud coverage in the images to 20% and to remove cloud shadows (cloud masking), resulting in a composite of images with minimal cloud and shadow coverage for the specified area (PPAs of the DF) for the selected periods (dry and rainy seasons).

The search for images, according to the specified filters, significantly reduced the time and effort associated with image pre-processing, which traditionally involves searching for images, downloading, removing clouds and shadows, and merging images (mosaic) to obtain a single image covering the entire study area. By using GEE, all these steps were incorporated into the same algorithm, allowing for the construction of a single image with all information extracted from aerial photography of different periods stored and processed in the cloud.

For this study, images were sought for the driest and wettest periods of the year. June to August was identified as the dry period, and November to January as the rainy period, as shown in Table 1. Initially, images from June to August 2015 were sought, but due to the satellite's launch in the same period (23 June 2015), there were no available images meeting the minimum requirements. Therefore, images from June to August 2016 and 2022 were sought. For the rainy period, images from

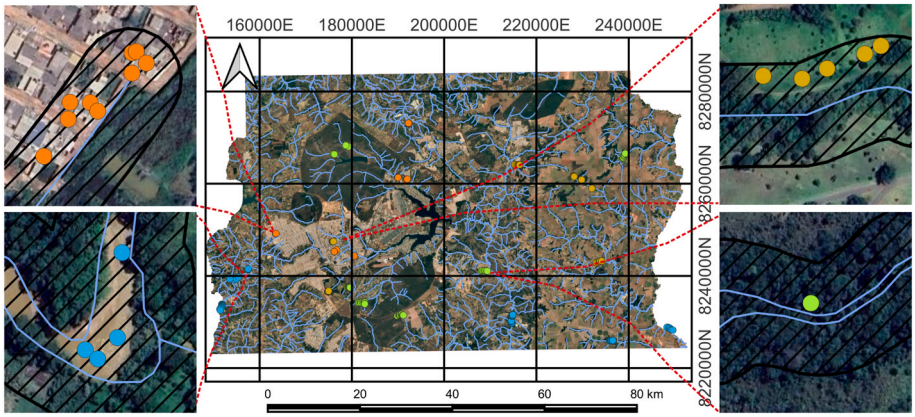

Fig. 3. Distribution of training and validation points.
Caption: Orange circles (top-left) represent the LC_Degradation: Urban areas**;** Blue circles (bottom-left) represent the LC_Water: Water bodies; Yellow circles (top-right) represent the LC_Agriculture: Agricultural areas; Green circles (bottom-right) represent the LC_Woody: Forest/native vegetation.

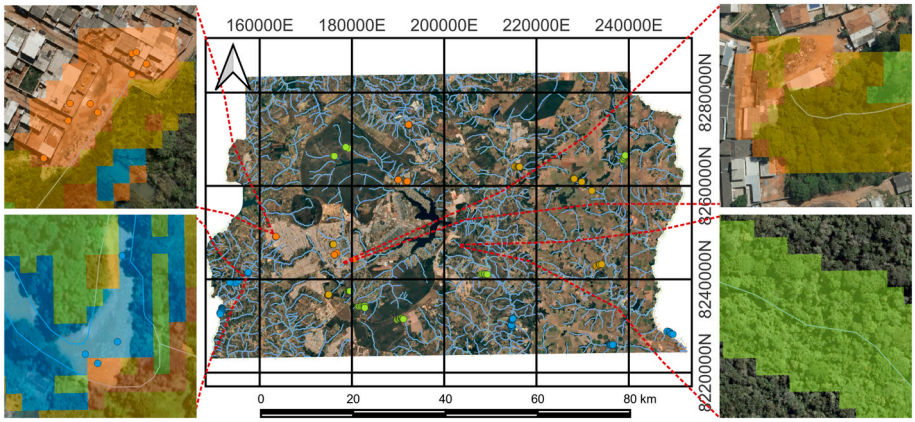

Fig. 4. Land use classification map for 2016.
Caption: Orange pixels represent the LC_Degradation: Urban areas**;** Blue pixels represent the LC_Water: Water bodies; Yellow pixels represent the LC_Agriculture: Agricultural areas; Green pixels represent the LC_Woody: Forest/native vegetation.

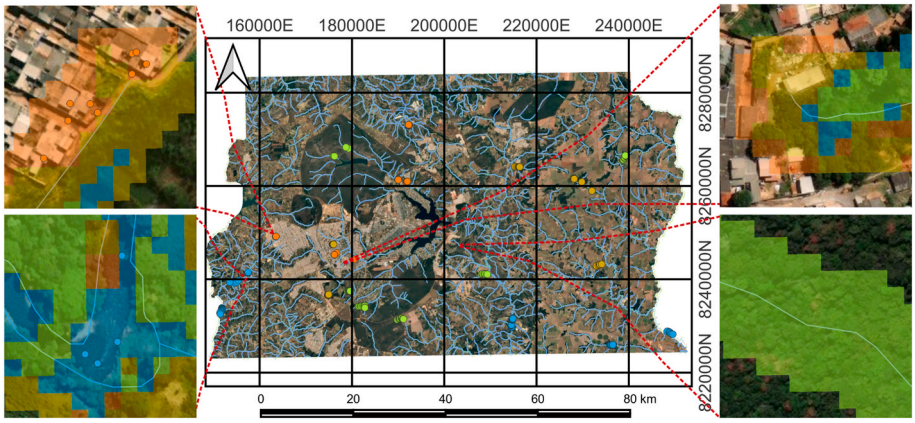

Fig. 5. Land use classification map for 2022.
Caption: Orange pixels represent the LC_Degradation: Urban areas**;** Blue pixels represent the LC_Water: Water bodies; Yellow pixels represent the LC_Agriculture: Agricultural areas; Green pixels represent the LC_Woody: Forest/native vegetation.

Table 1
– Average monthly precipitation for the Distrito Federal across two periods (1961–1990 and 1991–2020).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1961–1990	247	218	181	124	39	9	11	14	55	167	231	246
1991–2020	206	180	226	145	27	3	2	16	38	142	253	241

Source: INMET (2023).

November 2015 to January 2016, and from November 2022 to January 2023 were sought (INMET, 2023).

The dry months (June to August) yielded the best classification results, with overall accuracy between 86% and 88%, and a Kappa coefficient of 81% (2016) and 84% (2022). The rainy months (November to January) showed slightly lower results, with accuracy between 81% and 84%, and a Kappa coefficient between 73% and 79%, as shown in Table 2. The classification accuracy observed in this study aligns with findings from Phan et al. (2020) and Bessinger et al. (2022), reinforcing the reliability of RF in environmental monitoring. This consistency across diverse geographies highlights RF's robustness in adapting to varying ecological conditions. The Kappa coefficient (equation below) assesses the level of agreement between two sets of data, being a quantitative measure of the reliability of two evaluators judging the same dataset, corrected for the frequency with which evaluators present the same result.

$$k = \frac{Po - Pe}{1 - Pe}$$

Where Po is the observed agreement probability; Pe is the hypothetical

Table 2
– Overall accuracy and Kappa coefficient for land cover classification across different periods and seasons, 2015–2023.

	Jun/2016–Aug/2016	Jun/2022–Aug/2022	Nov/2015–Jan/2016	Nov/2016–Jan/2017	Nov/2022–Jan/2023
Overall accuracy	0.86	0.88	0.81	0.83	0.84
Kappa coefficient	0.81	0.84	0.73	0.76	0.79

expected probability.

According to Landis and Koch (1977), values above 81% present almost perfect agreement, and values between 61% and 80%, substantial agreement. The classification in the dry period was almost perfect and in the rainy period, substantial. Similar results were found by Phan et al. (2020), which obtained overall accuracy above 84.31%, and results between 77.66% and 89.90%, according to the different strategies for selecting the input images; and by Bessinger et al. (2022), which obtained an average overall accuracy of 82.28%, with values between 75.33% and 86.70%, and an average Kappa coefficient equal to 0.8068, with values between 0.7310 and 0.8550.

The reason for the difference observed in the Kappa coefficient between the rainy months (substantial agreement) and the dry months (almost perfect agreement) is likely due to cloud cover and shadows. Given the intensity of the rainy and dry seasons in the *Distrito Federal*, with monthly precipitation exceeding 200 mm between November and January, there's a high likelihood of encountering clouds (and shadows) in satellite images, even with specified filters, thus limiting the number of images and pixels available for classification. In contrast, during the dry period, with an average monthly precipitation of less than 15 mm, there's a greater probability of finding "cleaner" images with virtually no cloud presence.

Another possible reason for the observed classification difference is due to the change in pixel colouration of the LC_Water, LC_Degradation, and LC_Agriculture classes, as shown in the confusion matrices Table 3 and F1 score in Table 4. As can be observed in satellite images, river colouration tends towards brown hues during the rainy season, complicating classification since similar tones may be found in the LC_Agriculture and LC_Degradation classes, in areas of exposed soil. During the dry periods, rivers exhibit bluer tones, facilitating the differentiation between the LC_Water and LC_Agriculture classes.

The confusion matrix evaluates the performance of the classification carried out by a classification algorithm, presenting the distribution of records in their actual classes and their predicted classes. It offers a clearer understanding of what the classification model is getting right and the types of errors it is making. While the F1 score measures the model's accuracy by combining precision and recall scores extracted from the confusion matrix. The precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly. The recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive (Wagle et al., 2020).

Another challenge presented by the algorithm was the classification among LC_Agriculture, LC_Degradation, and LC_Woody during rainy periods, as there are green-coloured pixels in these three classes due to agricultural production, urban gardens, and the natural vegetation of the PPAs. In contrast, during the dry season, greener tones are usually exhibited by agriculture, owing to irrigation, while the other classes tend to show a drier appearance and more yellowish tones.

Producer's Accuracy measures the likelihood that a specific land use class within a study area is correctly classified, essentially indicating the frequency at which actual features on the ground are accurately depicted on the classified map. In opposition, User's Accuracy reveals how often a land use class identified on the classified map actually exists in reality. This distinction confirms some of the challenges the algorithm faces in classification during rainy periods, particularly for the LC_Agriculture class, which showed the lowest values, but also for other classes in comparison to the dry season, as shown in Table 4.

The choice of dry season results to meet the research objective—observing the preservation of the PPA in *Distrito Federal* (Brazil)—was based on accuracy results, the Kappa coefficient, confusion matrix, and F1 score findings, in addition to the reasons outlined above. The Figs. 4 and 5 above display the land use classification results of the PPAs, and Table 5 lists the total areas found for each class.

As observed in Figs. 4 and 5 and the areas detailed in Table 4, there

Table 4
– Producer and consumer accuracy and F1 Scores for land cover classification across different periods and seasons, 2015–2023.

Time Period	Category	Producer accuracy	Consumer accuracy	F1 Score
Jun/2016–Aug/2016	LC_WATER	0.85	0.85	0.85
	LC_DEGRADATION	0.85	0.89	0.87
	LC_WOODY	1.00	0.86	0.92
	LC_AGRICULTURE	0.77	0.83	0.80
Jun/2022–Aug/2022	LC_WATER	0.93	0.76	0.84
	LC_DEGRADATION	0.81	1.00	0.90
	LC_WOODY	0.87	0.95	0.91
	LC_AGRICULTURE	0.94	0.83	0.88
Nov/2015–Jan/2016	LC_WATER	0.93	1.00	0.96
	LC_DEGRADATION	0.75	0.69	0.72
	LC_WOODY	0.91	0.71	0.80
	LC_AGRICULTURE	0.40	1.00	0.57
Nov/2016–Jan/2017	LC_WATER	0.83	0.71	0.77
	LC_DEGRADATION	0.78	0.78	0.78
	LC_WOODY	0.95	0.95	0.95
	LC_AGRICULTURE	0.67	0.73	0.70
Nov/2022–Jan/2023	LC_WATER	0.86	0.92	0.89
	LC_DEGRADATION	0.76	0.84	0.80
	LC_WOODY	1.00	0.79	0.88
	LC_AGRICULTURE	0.79	0.85	0.81

Table 3
– Confusion matrix for land cover classification across different periods and seasons, 2015–2023.

Time Period	Category	LC_WATER	LC_DEGRADATION	LC_WOODY	LC_AGRICULTURE
Jun/2016–Aug/2016 - Dry	LC_WATER	17	2	1	0
	LC_DEGRADATION	1	17	0	2
	LC_WOODY	0	0	12	0
	LC_AGRICULTURE	2	0	1	10
Jun/2022–Aug/2022 - Dry	LC_WATER	13	0	1	0
	LC_DEGRADATION	0	13	0	3
	LC_WOODY	3	0	20	0
	LC_AGRICULTURE	1	0	0	15
Nov/2015–Jan/2016 - Rainy	LC_WATER	13	0	1	0
	LC_DEGRADATION	0	9	3	0
	LC_WOODY	0	1	10	0
	LC_AGRICULTURE	0	3	0	2
Nov/2016–Jan/2017 - Rainy	LC_WATER	5	0	0	1
	LC_DEGRADATION	2	14	0	2
	LC_WOODY	0	1	21	0
	LC_AGRICULTURE	0	3	1	8
Nov/2022–Jan/2023 - Rainy	LC_WATER	12	0	0	2
	LC_DEGRADATION	1	16	4	0
	LC_WOODY	0	0	15	0
	LC_AGRICULTURE	0	3	0	11

Table 5
– Land cover area proportions and total area across different periods and seasons, 2015–2023.

	Area (%)				
	Jun/2016–Aug/2016	Jun/2022–Aug/2022	Nov/2015–Jan/2016	Nov/2016–Jan/2017	Nov/2022–Jan/2023
LC_WATER	19.1%	13.2%	4.8%	6.3%	3.5%
LC_DEGRADATION	12.3%	11.7%	17.7%	6.3%	5.4%
LC_WOODY	50.6%	56.7%	59.2%	53.4%	60.6%
LC_AGRICULTURE	18.0%	18.4%	18.3%	33.9%	30.4%
Total Area	**164.87**	**184.28**	**103.52**	**151.9**	**164.87**

was an increase in the preserved area within the DF's PPAs (LC_Woody) between 2015 and 2023, growing by approximately 6% in both the dry and rainy seasons. This increase in vegetation area replaced spaces previously occupied by urban areas (LC_Degradation), a change that can be attributed to the DF's policies on the irregular occupation of PPAs lands and the spontaneous regeneration of vegetation once these areas are vacated. Another factor potentially explaining this shift could be the classification errors of the algorithm, as indicated in the confusion matrix between LC_Water and LC_Woody, and by the Producer's Accuracy, where the LC_Water class scored 0.85 in 2016, justifying an approximate 6% increase in one class and a corresponding decrease in the other.

Agricultural areas (LC_Agriculture) practically did not change, which can also be explained by the strong division between urban and agricultural areas, the first being allocated further west of the DF, and the second further east, in addition to the growth rates of the DF, which imposes extra pressure, limiting the growth of this class.

As for the total areas classified, as shown in Table 4, the total area for Nov/2015–Jan/2016 was the smallest of the five classifications carried out (103.5 km^2), while the others presented values above 150 km^2. The difference is due to the availability of images that meet the established minimum prerequisites, aiming to reduce clouds and shadows. However, probably because the Sentinel-2 satellite was launched in June 2015, its operation was probably not yet at full capacity during the period evaluated, which can be seen when comparing with other periods. Another factor that contributed to the difference was the period of drought and rain, with the classification in the dry period in 2022 being the one with the largest area, due to the low presence of clouds and, therefore, the greater availability of images that meet the minimum requirements.

In methodological terms, the decision to use high-resolution images from Sentinel-2 to monitor specific changes in vegetation within PPAs exemplifies a methodologically rigorous approach designed to capture detailed environmental changes over time. This is particularly critical in regions where minor alterations in land cover can have significant ecological impacts. Although some studies have also opted to use Sentinel-2 for its advantages Phan et al. (2020), there are indications of lower spatial consistency among the results, which could diminish its effectiveness in applications requiring high precision, such as detailed ecological monitoring or compliance with specific environmental regulations. In other words, our study not only focuses on achieving high classification accuracy (Radoux et al., 2016), but also contributes to methodological advancements in the field of environmental monitoring, with potential for broader applications in similar ecosystems globally.

The results demonstrated from the Kappa coefficient indicate a high degree of reliability and accuracy in the classification outcomes, essential for precise ecosystem monitoring and informed decision-making. While similar studies have achieved competent results, reporting an average overall accuracy of 82.28% with a range from 75.33% to 86.70% (Bessinger et al., 2022), the variability shown could undermine confidence in some of the classification outputs, particularly at the lower end of accuracy. Our study's robust model, combined with advanced machine learning techniques and comprehensive data analysis, not only enhances the granularity of environmental data analysis but also aligns with global efforts to meet Sustainable Development Goals through innovative technological applications. This alignment is not merely theoretical but demonstrated through practical application and quantifiable results, offering a significant contribution to both the academic community and practical field applications.

In conclusion, the application of the Random Forest algorithm through the Google Earth Engine platform has demonstrated significant potential in monitoring the preservation of PPAs in the *Distrito Federal* in alignment with SDG 6.6. The comparative analysis of satellite images from different periods has provided valuable insights into the changes in land use within PPAs, highlighting the effectiveness of machine learning techniques in environmental monitoring. The findings underscore the importance of high-resolution satellite imagery and advanced classification algorithms in enhancing the accuracy of land use classification, thereby contributing to the informed decision-making process for sustainable development. The results bridge existing gaps highlighted by Basu and Dasgupta (2021), providing a replicable and practical methodology for monitoring SDG 6.6. By leveraging accessible ML tools, it complements global research efforts, which advocate for the integration of advanced technologies in sustainable development practices (Yuan et al., 2020). This study not only reaffirms the critical role of technological advancements in achieving the SDGs but also sets a precedent for future research in the domain of environmental sustainability through the innovative use of machine learning.

4.1. Practical and managerial implications

The findings of this study offer significant practical and managerial implications for the monitoring and preservation of water-related ecosystems, essential for achieving SDG 6. Firstly, the successful application of the Random Forest method in classifying land use within PPA in the *Distrito Federal* (Brazil) underscores the effectiveness of statistical ML in environmental management. This approach demonstrates a replicable model for government and environmental agencies, suggesting that adopting advanced statistical ML techniques can enhance the accuracy and efficiency of ecosystem monitoring efforts.

The observed increase in native vegetation within PPAs, based on Sentinel-2 satellite imagery analysis, indicates not only the effectiveness of governmental conservation efforts but also highlights the importance of continuous and accurate monitoring to inform policy decisions. This increase also highlights the practical value of integrating ML tools into policy enforcement, aligning with findings on the need for robust regulatory frameworks (Ba et al., 2022). The use of freely available datasets and cloud-based platforms like GEE simplifies the data processing workflow, enabling more frequent evaluations of ecosystem preservation efforts without the need for extensive pre-processing or local data storage.

For managers and policymakers, these insights recommend a strategic focus on integrating ML tools into environmental monitoring frameworks to ensure data-driven decision-making. Additionally, the adaptability of the algorithm to various locales, contingent upon specific training, advises regional adaptations of these tools to account for ecosystem characteristics and seasonal variations.

A significant advancement of this study over similar approaches (Bessinger et al., 2022; Noi Phan et al., 2020; Pigola et al., 2021), is the intent to transcend more general applications that are not explicitly linked to any specific policy or global sustainability goals. Specifically,

this research targets the preservation of PPAs with a clear objective to achieve SDG 6, which pertains to clean water and sanitation. By monitoring land cover changes, the study demonstrates effective ecosystem management, directly supporting the preservation of water-related resources crucial for achieving SDG 6. These tools enable temporal assessments, ensuring accurate data for informed decision-making. Moreover, identifying degraded areas facilitates targeted interventions to mitigate erosion and runoff, safeguarding water bodies against pollution and sedimentation, thereby advancing clean water and sanitation objectives (Mustafa et al., 2022). This direct connection to the SDG targets adds a layer of political relevance and applicational significance to the research, enhancing its utility for both governmental and non-governmental organisations focused on meeting these global objectives.

Although similar studies have underscored important ecosystems, they tend to lack a prioritised focus on areas with immediate impact on human health and environmental sustainability (Bessinger et al., 2022). In this context, our choice to concentrate on PPAs, especially riparian zones critical for water quality and biodiversity, is particularly timely. This focused approach aligns closely with specific environmental and conservation goals, making the research highly relevant for conservation efforts and policymaking, particularly in achieving SDG 6. Overall, our work, by employing advanced machine learning integrated with high-resolution satellite imagery, contributes to methodological advancements in the field and sets objective guidelines for practical implementation or scalability. In doing so, it adds to the existing empirical literature (Nilashi et al., 2023; Pigola et al., 2021; Requejo-Castro et al., 2020; Jiang et al., 2024; Naghibi et al., 2016), which often lacks specific examples of implementation as well as the outcomes such implementations could yield.

In summary, this study advocates for the broader adoption of ML technologies in environmental policy and management practices. It highlights the potential of such technologies to support the global endeavour of sustainable development and environmental stewardship, offering a foundation for informed public policy formulation and conservation efforts.

Future research directions should include exploring hybrid models combining RF and other statistical machine-learning algorithms, such as classification or clustering. Classification algorithms, such as RF, allocate objects (pixels) into classes or groups based on input variables, and response (or output) variables are required to train the model. This model can then be applied to test data sets that contain only the input variables. Al-Obeidat et al. (2015) compared the performance of three classification algorithms: (i) Decision Tree C4.5, (ii) Decision Tree ID3, and (iii) a hybrid model with the combination of Multi-Criteria Decision Analysis (MCDA) and Decision Tree algorithms. The overall accuracy found was 89% (kappa = 0.8829) using the hybrid model, 86% (kappa = 0.8572) using Decision Tree C4.5, and 82% (kappa = 0.82) using Decision Tree ID3.

Integrating clustering algorithms would further refine classification accuracy, thereby broadening the utility of these methods across different ecosystems and geographical areas. Clustering algorithms are an unsupervised learning method that does not require a training data set. The algorithm seeks to understand the data independently, grouping the objects according to a similarity measure. In this case, tests could be performed, specifying the number of clusters (the same four used in this research) or letting the algorithm estimate the number of groups as part of the analysis. Then, apply the RF algorithm (or another classification algorithm), comparing whether the results were better when the clusters were selected by the clustering algorithm, instead of allocating the validation points, as presented in this paper. Incorporating spatial validation methods could enhance model generalisability by mitigating overfitting risks inherent in environmental classification tasks (Meyer et al., 2019).

Techniques such as K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models provide diverse methods for efficiently grouping data points. The integration of clustering algorithms with classification tasks can improve accuracy and offer deeper insights into the relationships within the data, presenting a promising area for future research exploration.

4.2. Social and economic implications

The findings from this research on the application of statistical ML, particularly the Random Forest method, to monitor and preserve water-related ecosystems, carry profound social and economic implications. Firstly, the demonstrated increase in native vegetation within the PPA of the *Distrito Federal* (Brazil) highlights the tangible benefits of employing advanced ML techniques in environmental conservation efforts. This growth not only signifies the recovery and preservation of ecosystems but also suggests a positive trajectory towards achieving SDG 6, which is vital for enhancing human well-being, driving economic and social progress, and preserving ecosystems crucial to water security.

From a social perspective, the improvement in ecosystem preservation directly contributes to the well-being of communities by ensuring access to clean water and sanitation facilities. It reflects an advancement in public health standards and supports the broader objective of eradicating poverty (SDG 1) and enhancing quality education (SDG 4), as healthier ecosystems are foundational to social development and economic prosperity.

Economically, the application of ML in environmental monitoring can lead to more efficient use of resources, reducing the costs associated with traditional data collection and analysis methods. The use of freely available datasets and cloud-based platforms like GEE illustrates the potential for cost-effective environmental management practices that can be adopted by governmental and environmental agencies. Leveraging increasing amounts of electronic data through such platforms enhances regulatory effectiveness and supports efforts to mitigate environmental harms (Hino et al., 2018). This approach enables a more dynamic allocation of resources towards areas of critical need, enhancing the effectiveness of conservation efforts and potentially leading to economic savings.

Moreover, the study's methodology, offering a replicable model for ecosystem monitoring, underscores the importance of technology and innovation in addressing global sustainability challenges. The accessibility of advanced ML techniques promises to democratise the monitoring of SDGs, enabling developing countries to participate more actively in global sustainability efforts. This can lead to economic development opportunities, fostering global partnerships and collaboration towards achieving the SDGs (Hofmann, 2021; Quinlivan et al., 2020).

In summary, the social and economic implications of this study advocate for the integration of machine learning technologies in environmental policy and management, highlighting their potential to contribute significantly to sustainable development, economic efficiency, and social well-being. Future research should continue to explore and refine these methodologies, expanding their application across diverse ecosystems and geographical areas to maximize their global impact.

5. Conclusions

This study has illuminated the potential of statistical ML as a robust tool in supporting the monitoring of the Sustainable Development Goals (SDGs), particularly SDG 6.6, which aims to protect and restore water-related ecosystems. By employing classification algorithms, notably the Random Forest method, this research has monitored the preservation of the *Distrito Federal*'s primary water-related ecosystem, the PPAs, through Sentinel-2 satellite imagery analysis across two periods: the inception of the SDGs in 2015–2016 and a more recent evaluation in 2022.

The observed increase in native vegetation within the PPAs by

approximately 6% from 2015 to 2022 signifies progress in biodiversity conservation, as they are home to a diverse range of flora and fauna. This growth plays a crucial role in protecting water bodies and preventing erosion, as PPAs function as ecological corridors, filtering sediments and pollutants, which helps maintain both water quality and quantity. Furthermore, this increase underscores the effectiveness of government conservation efforts and highlights the uncertainties in the algorithm's classification, as demonstrated by confusion matrices and accuracy metrics.The study's findings, demonstrating higher accuracy during dry periods with a near-perfect Kappa coefficient, align with existing literature on statistical ML classification methods. Utilizing free datasets, open-source remote sensing software, and cloud-based platforms like GEE has streamlined the data processing workflow, eliminating the need for extensive pre-processing and local data storage. This methodology not only facilitates rapid data analysis but also underscores the accessibility and technological innovation inherent in the GEE platform, enabling SDG monitoring in developing countries and enhancing the frequency of such evaluations globally.

Furthermore, the adaptability of the algorithm to various locales, contingent upon model training for specific areas, suggests a broader applicability for monitoring PPAs beyond the DF, albeit with considerations for regional ecosystem characteristics and seasonal variations. The study's approach, leveraging cloud-based platforms and machine learning, offers a model for environmental preservation and public policy formulation, providing updated data for governmental and environmental agencies to inform decision-making and conservation efforts. Future research should explore the integration of clustering algorithms to refine classification accuracy and expand the algorithm's utility across diverse ecosystems and geographical contexts, thereby contributing to the global endeavour of sustainable development and environmental stewardship.

CRediT authorship contribution statement

Murilo de Carvalho Marques: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Abdoulaye Aboubacari Mohamed:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Paulo Feitosa:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paulo Feitosa reports was provided by University of São Paulo. Paulo Feitosa reports a relationship with University of São Paulo that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., Alyaman, M., 2021. Poverty classification using machine learning: the case of Jordan. Sustain. Times 13, 1–16. https://doi.org/10.3390/su13031412.

Al-Obeidat, F., Al-Taani, A., Belacel, N., Feltrin, L., Banerjee, N., 2015. A fuzzy decision tree for processing satellite images and landsat data. Procedia Comput. Sci. 52, 1192–1197.

Arora, N.K., Mishra, I., 2022. Sustainable development goal 6: global water security. Environ. Sustain. 5, 271–275. https://doi.org/10.1007/s42398-022-00246-5.

Ba, S., Onyeabor, E.U., Moneke, A.N., 2022. The current legal framework for pollution control in the Niger River Basin relative to SDG 6.3. Water Int. 47, 1217–1234. https://doi.org/10.1080/02508060.2022.2073756.

Basu, M., Dasgupta, R., 2021. Where do we stand now? A bibliometric analysis of water research in support of the sustainable development goal 6. Water (Switzerland) 13, 1–18. https://doi.org/10.3390/w13243591.

Bebbington, J., Unerman, J., 2018. Achieving the united nations sustainable development goals: an enabling role for accounting research. Account Audit. Account. J. 31, 2–24. https://doi.org/10.1108/AAAJ-05-2017-2929.

Bessinger, M., Lück-Vogel, M., Skowno, A., Conrad, F., 2022. Landsat-8 based coastal ecosystem mapping in South Africa using random forest classification in Google Earth Engine. South Afr. J. Bot. 150, 928–939. https://doi.org/10.1016/j.sajb.2022.08.014.

Bhaduri, A., Bogardi, J., Siddiqi, A., Voigt, H., Vörösmarty, C., Pahl-Wostl, C., Bunn, S.E., Shrivastava, P., Lawford, R., Foster, S., Kremer, H., Renaud, F.G., Bruns, A., Osuna, V.R., 2016. Achieving sustainable development goals from a water perspective. Front. Environ. Sci. 4. https://doi.org/10.3389/fenvs.2016.00064.

Chen, Q., Liu, Z., 2019. How does openness to innovation drive organizational ambidexterity? the mediating role of organizational learning goal orientation. IEEE Trans. Eng. Manag. 66, 156–169. https://doi.org/10.1109/TEM.2018.2834505.

Congalton, R.G., Green, K., 2019. Assessing the Accuracy of Remotely Sensed Data : Principles and Practices, Third Edition, Assessing the Accuracy of Remotely Sensed Data. CRC Press. https://doi.org/10.1201/9780429052729.

Denu, M.K., Bentley, Y., Duan, Y., 2023. Social sustainability performance: developing and validating measures in the context of emerging African economies. J. Clean. Prod. 412, 137391. https://doi.org/10.1016/j.jclepro.2023.137391.

Diep, L., Martins, F.P., Campos, L.C., Hofmann, P., Tomei, J., Lakhanpaul, M., Parikh, P., 2021. Linkages between sanitation and the sustainable development goals: a case study of Brazil. Sustain. Dev. 29, 339–352. https://doi.org/10.1002/sd.2149.

European Commission, 2021. A Green and Digital Transformation of the EU. A Green Digit. EU.

Fisher, J., Allen, S., Yetman, G., Pistolesi, L., 2024. Assessing the influence of landscape conservation and protected areas on social wellbeing using random forest machine learning. Sci. Rep. 14 (1), 11357.

Fu, B., Wang, S., Zhang, J., Hou, Z., Li, J., 2019. Unravelling the complexity in achieving the 17 sustainable-development goals. Natl. Sci. Rev. 6, 382–383. https://doi.org/10.1093/nsr/nwz029.

Fuente, D., Allaire, M., Jeuland, M., Whittington, D., 2020. Forecasts of mortality and economic losses from poor water and sanitation in sub-Saharan Africa. PLoS One 15. https://doi.org/10.1371/journal.pone.0227611.

Guo, H., Dou, C., Chen, H., Liu, J., Fu, B., Li, X., Zou, Z., Liang, D., 2023. SDGSAT-1: the world's first scientific satellite for sustainable development goals. Sci. Bull. 68, 34–38. https://doi.org/10.1016/j.scib.2022.12.014.

Guppy, L., Mehta, P., Qadir, M., 2019. Sustainable development goal 6: two gaps in the race for indicators. Sustain. Sci. 14, 501–513. https://doi.org/10.1007/s11625-018-0649-z.

Heras, A.D. Las, Luque-Sendra, A., Zamora-Polo, F., 2020. Machine learning technologies for sustainability in smart cities in the post-covid era. Sustain. Times 12, 1–25. https://doi.org/10.3390/su12229320.

Herrera, V., 2019. Reconciling global aspirations and local realities: challenges facing the Sustainable Development Goals for water and sanitation. World Dev. 118, 106–117. https://doi.org/10.1016/j.worlddev.2019.02.009.

Hino, M., Benami, E., Brooks, N., 2018. Machine learning for environmental monitoring. Nat. Sustain. 1 (10), 583–588.

Hofmann, P., 2021. Meeting WASH SDG6: insights from everyday practices in Dar es Salaam. Environ. Urbanization 33, 173–192. https://doi.org/10.1177/0956247820957280.

IBGE, 2023. Malha Municipal [WWW Document]. URL. https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15774-malhas.html (accessed 8.8.23).

INMET, 2023. Gráficos Climatológicos [WWW Document]. URL. https://clima.inmet.gov.br/GraficosClimatologicos/DF/83377 (accessed 8.18.23).

Jain, A., Gue, I.H., Jain, P., 2023. Research trends, themes, and insights on artificial neural networks for smart cities towards SDG-11. J. Clean. Prod. 412, 137300. https://doi.org/10.1016/j.jclepro.2023.137300.

Jiang, H., Gu, J., Xi, H., Yu, Q., Wang, X., Liu, Y., 2024. Machine learning clustering algorithm for water environmental monitoring. Int. J. Pattern Recogn. Artif. Intell.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159. https://doi.org/10.2307/2529310.

Lin, J., Qiu, S., Tan, X., Zhuang, Y., 2023. Measuring the relationship between morphological spatial pattern of green space and urban heat island using machine learning methods. Build. Environ. 228, 109910. https://doi.org/10.1016/j.buildenv.2022.109910.

Madrazo-Ortega, D., Molinos-Senante, M., 2023. Quantifying progress made in achieving sustainable development goal 6 in Chile: a holistic and local approach. Sustain. Times 15. https://doi.org/10.3390/su15054125.

Martínez-Córdoba, P.J., Raimo, N., Vitolla, F., Benito, B., 2020. Achieving sustainable development goals. Efficiency in the Spanish clean water and sanitation sector. Sustain. Times 12, 1–13. https://doi.org/10.3390/su12073015.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinf. 10, 1–16. https://doi.org/10.1186/1471-2105-10-213.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815.
Miao, J., Song, X., Zhong, F., Huang, C., 2023. Sustainable development goal 6 assessment and attribution analysis of underdeveloped small regions using integrated multisource data. Rem. Sens. 15. https://doi.org/10.3390/rs15153885.
Mustafa, S., Jamil, K., Zhang, L., Girmay, M.B., 2022. Does public awareness matter to achieve the UN's sustainable development goal 6: clean water for everyone? J. Environ. Publ. Health. https://doi.org/10.1155/2022/8445890.
Naghibi, S.A., Pourghasemi, H.R., Dixon, B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188, 1–27.
Nhamo, G., Nhemachena, C., Nhamo, S., 2019. Is 2030 too soon for Africa to achieve the water and sanitation sustainable development goal? Sci. Total Environ. 669, 129–139. https://doi.org/10.1016/j.scitotenv.2019.03.109.
Nilashi, M., Keng Boon, O., Tan, G., Lin, B., Abumalloh, R., 2023. Critical data challenges in measuring the performance of sustainable development goals: solutions and the role of big-data analytics. Harvard Data Sci. Rev. 5. https://doi.org/10.1162/99608f92.545db2cf.
Nkiaka, E., Bryant, R.G., Okumah, M., Gomo, F.F., 2021. Water security in sub-Saharan Africa: understanding the status of sustainable development goal 6. Wiley Interdiscip. Rev. Water 8, 1–20. https://doi.org/10.1002/wat2.1552.
Noi Phan, T., Kuch, V., Lehnert, L.W., 2020. Land cover classification using google earth engine and random forest classifier-the role of image composition. Rem. Sens. 12. https://doi.org/10.3390/RS12152411.
Otukei, J.R., Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. Int. J. Appl. Earth Obs. Geoinf. 12, 527–531.
Partzsch, L., Hartung, K., Lümmen, J., Zickgraf, C., 2021. Water in your coffee? Accelerating SDG 6 through voluntary certification programs. J. Clean. Prod. 324, 129252. https://doi.org/10.1016/j.jclepro.2021.129252.
Pereira, M.A., Marques, R.C., 2021. Sustainable water and sanitation for all: are we there yet? Water Res. 207, 117765. https://doi.org/10.1016/j.watres.2021.117765.
Phan, T.N., Kuch, V., Lehnert, L.W., 2020. Land cover classification using Google Earth Engine and random forest classifier – the role of image composition. Rem. Sens. 12, 2411.
Pigola, A., da Costa, P.R., Carvalho, L.C., da Silva, L.F., Kniess, C.T., Maccari, E.A., 2021. Artificial intelligence-driven digital technologies to the implementation of the sustainable development goals: a perspective from Brazil and Portugal. Sustain. Times 13. https://doi.org/10.3390/su132413669.
Porciello, J., Ivanina, M., Islam, M., Einarson, S., Hirsh, H., 2020. Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. Nat. Mach. Intell. 2, 559–565. https://doi.org/10.1038/s42256-020-00235-5.
Quinlivan, L., Chapman, D.V., Sullivan, T., 2020. Validating citizen science monitoring of ambient water quality for the United Nations sustainable development goals. Sci. Total Environ. 699, 134255. https://doi.org/10.1016/j.scitotenv.2019.134255.
Radoux, J., Chomé, G., Jacques, D.C., Waldner, F., Bellemans, N., Matton, N., Lamarche, C., D'Andrimont, R., Defourny, P., 2016. Sentinel-2's potential for sub-pixel landscape feature detection. Rem. Sens. 8. https://doi.org/10.3390/rs8060488.
Ramos, S.B., de Paula Silva, J., Bolela, C.A., de Andrade, M., 2018. Prediction of human development from environmental indicators. Soc. Indicat. Res. 138, 467–477. https://doi.org/10.1007/s11205-017-1693-2.
Requejo-Castro, D., Giné-Garriga, R., Pérez-Foguet, A., 2020. Data-driven Bayesian network modelling to explore the relationships between SDG 6 and the 2030 Agenda. Sci. Total Environ. 710, 136014. https://doi.org/10.1016/j.scitotenv.2019.136014.
Robins, L., Burt, T.P., Bracken, L.J., Boardman, J., Thompson, D.B.A., 2017. Making water policy work in the United Kingdom: a case study of practical approaches to strengthening complex, multi-tiered systems of water governance. Environ. Sci. Pol. 71, 41–55. https://doi.org/10.1016/j.envsci.2017.01.008.
Sheffield, J., Wood, E.F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., Verbist, K., 2018. Satellite remote sensing for water resources management: potential for supporting sustainable development in data-poor regions. Water Resour. Res. 54, 9724–9758. https://doi.org/10.1029/2017WR022437.
Singpai, B., Wu, D., 2020. Using a DEA–AutoML approach to track SDG achievements. Sustain. Times 12, 1–26. https://doi.org/10.3390/su122310124.
SISDIA, 2023. Hidrografia CRH [WWW Document]. URL. https://sisdia.df.gov.br/portal/home/item.html?id=3e890dd9c6c24cfba6808722628f0c7 (accessed 8.8.23).
Taka, M., Ahopelto, L., Fallon, A., Heino, M., Kallio, M., Kinnunen, P., Niva, V., Varis, O., 2021. The potential of water security in leveraging Agenda 2030. One Earth 4, 258–268. https://doi.org/10.1016/j.oneear.2021.01.007.
Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google Earth Engine for geo-big data applications: a meta-analysis and systematic review. ISPRS J. Photogrammetry Remote Sens. 164, 152–170.
United Nations, 2022. The Sustainable Development Goals Report 2019, the Sustainable Development Goals Report.
Vazquez-Brust, D., Piao, R.S., de Melo, M.F. de S., Yaryd, R.T., Carvalho, M., 2020. The governance of collaboration for sustainable development: exploring the "black box.". J. Clean. Prod. 256, 120260. https://doi.org/10.1016/j.jclepro.2020.120260.
Wagle, N., Acharya, T., Kolluru, V., Huang, H., Lee, D., 2020. Multi-temporal land cover change mapping using google earth engine and ensemble learning methods. Appl. Sci. 10, 8083.
Wang, D., Fu, J., Xie, X., Ding, F., Jiang, D., 2022. Spatiotemporal evolution of urban-agricultural-ecological space in China and its driving mechanism. J. Clean. Prod. 371, 133684. https://doi.org/10.1016/j.jclepro.2022.133684.
Yamane, T., Kaneko, S., 2021. Is the younger generation a driving force toward achieving the sustainable development goals? Survey experiments. J. Clean. Prod. 292, 125932. https://doi.org/10.1016/j.jclepro.2021.125932.
Yao, T., Li, J., 2023. Environmental sustainability performance assessment in relation to visibility in African regions with interpretable machine learning. J. Clean. Prod. 428, 139414. https://doi.org/10.1016/j.jclepro.2023.139414.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., et al., 2020. Deep learning in environmental remote sensing: achievements and challenges. Rem. Sens. Environ. 241, 111716.
Zhang, K., Qing, Y., Umer, Q., Asmi, F., 2023. How construction and demolition waste management has addressed sustainable development goals: exploring academic and industrial trends. J. Environ. Manag. 345, 118823. https://doi.org/10.1016/j.jenvman.2023.118823.
Zhang, W., Muhammad, W., Supunsala, S., Alessi, D.S., Tack, F.M.G., Sik, Y., 2023. Science of the Total Environment Machine learning based prediction and experimental validation of arsenite and arsenate sorption on biochars. Sci. Total Environ. 904, 166678. https://doi.org/10.1016/j.scitotenv.2023.166678.
Zhu, S., Preuss, N., You, F., 2023. Advancing sustainable development goals with machine learning and optimization for wet waste biomass to renewable energy conversion. J. Clean. Prod. 422, 138606. https://doi.org/10.1016/j.jclepro.2023.138606.