Caatinga - Appendix

MapBiomas 10-meters

Algorithm Theoretical Basis Document (ATBD)

Collection 3.0

Version 1

General Coordinator
Washington de Jesus Sant'anna da Franca Rocha (UEFS)

Team
Diego Pereira Costa (GEODATIN/UEFS)
Rodrigo Nogueira de Vasconcelos (GEODATIN/UEFS)
Nerivaldo Afonso Santos (GEODATIN/UEFS)
Rafael Oliveira Franca Rocha (GEODATIN/UEFS)
Soltan Galano Duverger (GEODATIN/UEFS)
Deorgia Tayane Mendes de Souza (UEFS/PPGM)
Jocimara Souza Lobão (UEFS/PPGM)

April, 2026

1. OVERVIEW OF THE CAATINGA CLASSIFICATION METHOD

This document outlines the specific methodologies employed to create annual land use and land cover (LULC) maps of the Caatinga biome as part of the MapBiomas 10-meters initiative. These products are based on Sentinel-2 high spatial remote sensing data and fully implemented into Google Earth Engine (GEE) environment.

Aligning with the methodology for 30-meter Landsat LULC products, the 10-meter datasets transitioned from the Random Forest (RF) model used in previous collections to Gradient Tree Boosting (GTB). Due to its superior performance, GTB was adopted as the primary classifier for all biome products.

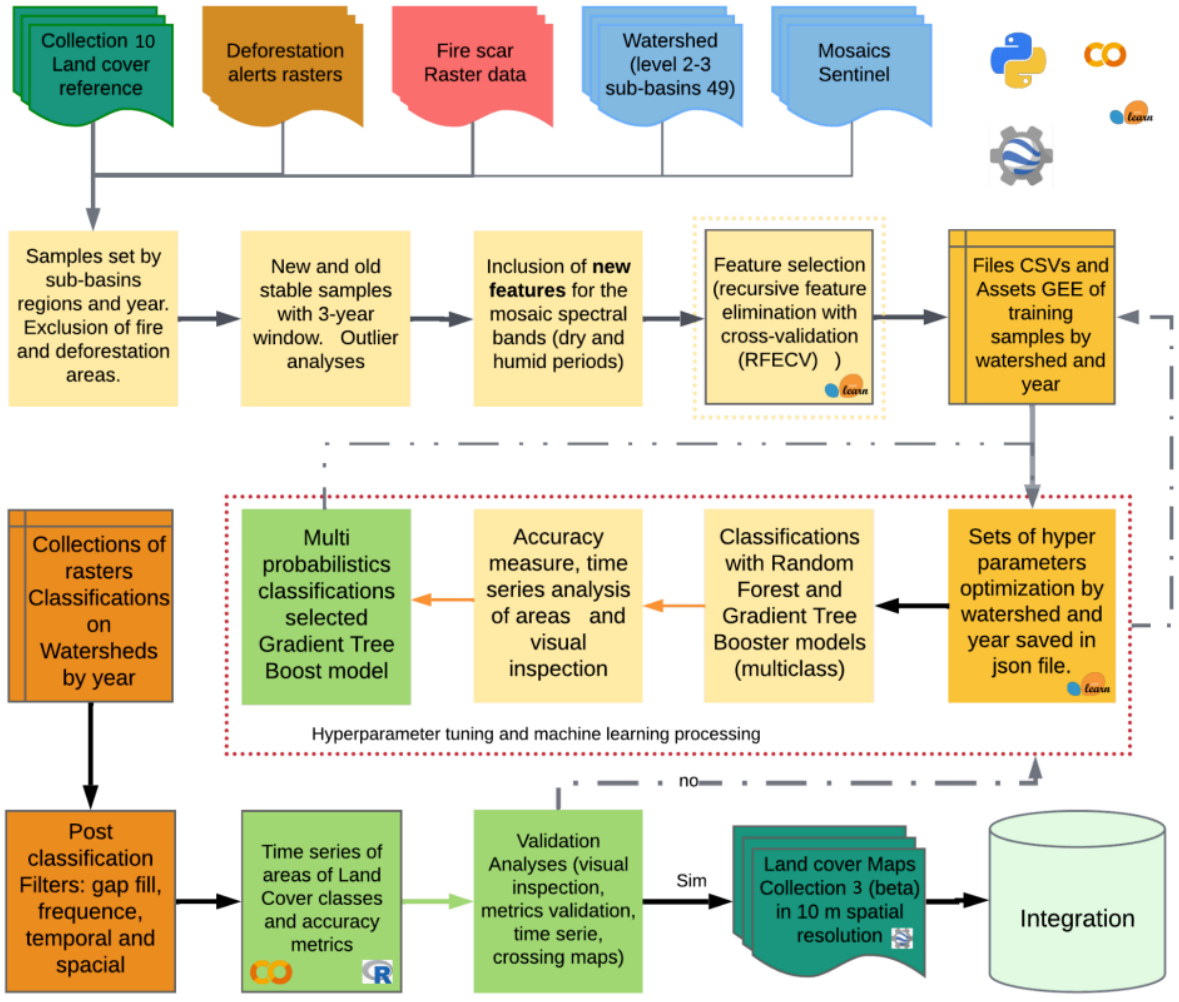

Figure 1. Detailed steps of MapBiomas 10 m Collection 3 (beta) (2017-2024) in the Caatinga biome.

As a vital cross-cutting dataset, the *Photovoltaic Solar Plants* layer utilizes a U-Net deep learning architecture first debuted in LULC Collection 10 and subsequently adopted for MapBiomas 10-meter Collection 3. This layer maps the entirety of Brazil's registered photovoltaic infrastructure using ANEEL data. For model training, the team utilized precise plant coordinates, stored as GEE assets, to generate the necessary image "patches" for the deep learning workflow.

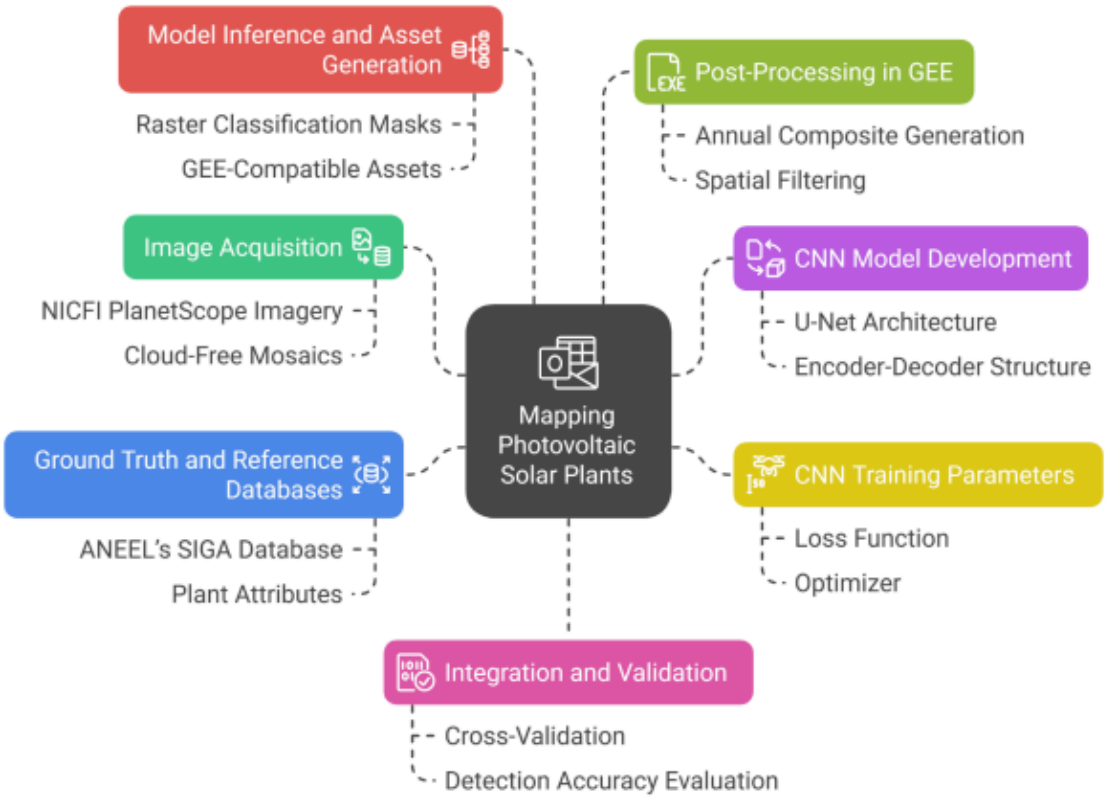

Figure 2. Processing workflow. All processing steps are represented using distinct colors to illustrate each stage of the methodology.

The MapBiomas Caatinga team, in alignment with other project teams, employed a stratified sampling strategy, segmenting the biome into base regions. To enhance the reliability of automatically collected samples, a mask was generated to exclude areas affected by temporal phenomena, such as wildfires and deforestation. Subsequently to the mask application, sample collection was restricted to those pixels maintaining consistent classification in the pre- and post-collection periods, and selecting information from the mosaic bands only within the mask-permitted areas. The resulting samples underwent a feature selection process, wherein spectral bands that optimized classifier performance in land cover class prediction were identified.

The feature selection process utilized the Recursive Feature Elimination with Cross-Validation (RFECV) methodology. RFECV's goal is to select the most important features, those that contribute the most to the model's performance. The recursive process of RFECV started with all features, training of the model, ranking features by importance, removing the least important, and repeated all steps with cross-validation. The cross-validation part was crucial because it helped in determining the right number of features without overfitting (Chen and Guestrin, 2016).

To optimize the classifier's parameters for maximum accuracy, the samples containing the information from the columns or bands selected in the previous process were used as a basis for hyperparameter tuning, aiming to determine the optimal parameters for each classification region. The method employed was Grid Search, a classic hyperparameter tuning technique used to identify the ideal combination of machine learning model hyperparameters. This technique evaluated the model's performance for each hyperparameter combination, identifying the one with the best performance (e.g., highest accuracy, lowest error). The combination with cross-validation enhanced the robustness of the results, as demonstrated by Bergstra and Bengio (2012).

Gradient Tree Boosting was employed as the classification algorithm. It is an ensemble learning method that sequentially aggregates multiple weak decision trees, typically weak, where each subsequent tree corrects the errors of its predecessors (Friedman, 2001; Natekin and Knoll, 2013; Abdi, 2020; Ou et al., 2023). The algorithm takes advantage of the gradient method and the derivatives of the loss function to perform an iterative optimization of the model. This classifier was selected over Random Forest for the Caatinga biome due to its superior performance, attributed to specific characteristics relevant to the region:

● Sequential Learning: Each tree is trained to minimize the residual (error) of the previous tree, progressively refining predictions.

● Customizable Loss Function: It can be adapted to regression, binary classification, or multiclass classification problems.

● Regularization: Hyperparameters such as learning_rate, max_depth, and subsample prevent overfitting by controlling model complexity.

● Feature Importance: Identifies which input variables (e.g., spectral bands, vegetation indices) most significantly influence predictions.

The Remote sensing data have complex characteristics that align well with Gradient Boosting strengths:

1. Nonlinearity: Captures intricate relationships between spectral bands and target variables (e.g., land cover classes).

2. High Dimensionality: Effectively handles datasets with many features, such as spectral bands, indices, and texture metrics.

3. Imbalanced Data: Addresses scenarios like detecting deforestation in small areas versus intact forests by weighting classes or using sampling techniques.

Gradient Tree Boosting is a powerful tool for remote sensing due to its flexibility, robustness to noise, and capacity to handle complex datasets. By combining sequential error correction with regularization techniques, it delivers high accuracy in tasks ranging from land cover mapping to environmental monitoring.

2. METHODOLOGY EVOLUTION AND CLASSES MAPPED

Annual mosaics were classified for the period from 2017 to 2024, encompassing ten land cover and land use (LCLU) classes: Forest Formation, Savanna Formation, Grassland, Herbaceous Sandbank vegetation, Mosaic of Uses, Non-Vegetated Areas, Other Non-Vegetated Areas, Rocky Outcrops, Photovoltaic Solar Plants and Water, aligning with MapBiomas Collection 10.

The evolution of mapping methodologies across successive collections is summarized in Table 1, while Table 2 provides a comprehensive description of the LULC classes identified. Several of these classes were refined through integration with cross-cutting themes. (Classes full description available at: Legend Description).

Table 1.Overview of LULC collections of the Caatinga biome.

Collection	Time Interval	Method	Class	Key Improvements
1	2016-2022	RF/GTB	-Forest Formation -Savanna Formation -Grassland -Mosaic of Uses -Water -Non-vegetated Areas -Other Non-vegetated Areas -Rocky Outcrops -Herbaceous Sandbank Vegetation.	
2	2016-2023	GTB/cluster	-Forest Formation -Savanna Formation -Grassland -Mosaic of Uses -Water -Non-vegetated Areas -Other Non-vegetated Areas -Rocky Outcrops -Herbaceous Sandbank Vegetation.	*Rocky outcrops* class was made using a cluster model
3	2017-2024	GTB/cluster/ U-net	-Forest Formation -Savanna Formation -Grassland -Mosaic of Uses -Water -Non-vegetated Areas -Other Non-vegetated Areas -Rocky Outcrops -Herbaceous Sandbank Vegetation. -Photovoltaic Solar Plants	Use of Google Satellite Embedding into Feature Space Photovoltaic Solar Plants class was mapped using the U-Net model.

Table 2. Land cover and land use classes considered for digital classification of Landsat mosaics in the Caatinga biome in the MapBiomas Collection 10.

Legend class	ID	Natural / Anthropic	Land cover / Land use	General description
1.1 Forest Formation	3	Natural	Land cover	Vegetation with predominance of continuous canopy-Savana- Estépica, Florestalada, Seasonal Semi-Deciduous and Deciduous Forest.
1.2 Savanna Formation	4	Natural	Land cover	Vegetation with predominance of semi-continuous canopy species - savanna-shrub savanna- savanna woodland.
1.4 Herbaceous Sandbank Vegetation	49	Natural	Land cover	Herbaceous Sandbank Vegetation includes herbaceous plant communities dominated by shrubs or small trees. These species are

				frequently wide-spread and occur in coastal areas of Southeastern Brazil
2.2 Grassland	12	Natural	Land cover	Vegetation with predominance of herbaceous species (steppe Savannah Grassy-Woody, Savanna park, Savanna Grassy-Woody.
2.4 Rocky Outcrop	29	Natural	Land cover	Rocks naturally exposed on the earth's surface without soil cover, often with the partial presence of rupicolous vegetation and high slope.
3.3 Mosaic of Uses	21	Anthropic	Land use	Use agriculture areas where it was not possible to distinguish between pasture and agriculture.
4. Non vegetated Area	22	Anthropic	Land use	Beach and Dune, Urban Infrastructure and Mining.
4.4. Other non Vegetated Areas	25	Anthropic	Land cover	Non-permeable surface areas (infrastructure, urban expansion or mining) not mapped into their classes and regions of exposed soil in natural or crop areas. Mixed class that includes natural and anthropic areas.
4.6 Photovoltaic power plant	75	Anthropic	Land cover	A "photovoltaic power plant" is a medium to large-scale installation designed to generate electricity directly from sunlight, primarily focused on commercializing the energy. In Brazil, plants with a capacity greater than 5 MW are considered large-scale, while those up to 5 MW are classified as mini-generation, according to regulations (Law 14.182/2021; Law 10.438/2002; Decree 5.025/2004; ANEEL Resolution 127/2004). The electricity generated is connected to the National Interconnected System (SIN), which distributes power throughout the country. In terms of land use, these plants occupy significant areas: it's estimated that an installation in tropical regions requires about 1 ha per MW using fixed modules, potentially varying to 2–3 ha/MW depending on technology (trackers) and panel arrangement. National examples confirm this range: the Nova Olinda Solar Park (292 MW across 690 ha ≈ 2.4 ha/MW), and the Pirapora Solar Complex (321 MW across ≈ 1,500 ha, about 4.7 ha/MW).
5. Water	33	Natural / Anthropic	Land cover / Land use	Rivers, lakes, dams, reservoir and other water bodies

The following sections provide a comprehensive account of the Collection 3 methodology, detailing each developmental stage and implementation phase while emphasizing the strategic enhancements integrated into the classification workflow.

Methodologies employed in previous collections are accessible through the MapBiomas ATBD link (https://mapbiomas.org/download-dos-atbds).

3. IMAGE MOSAICS

The Collection 3 LULC classification for the Caatinga biome involves the generation of annual Sentinel-2 image mosaics. For the MapBiomas 10-meters initiative, the construction strategy leverages methodological advancements established in previous Landsat-based collections (such as Collection 10). In Collections 1 and 2, annual mosaics were derived exclusively from Sentinel-2 surface reflectance imagery (COPERNICUS/S2_SR_HARMONIZED). The process utilizes the full spectral suite available, including blue, green, red, red-edge (1, 2, 3, and 4), near-infrared (NIR), and shortwave infrared (SWIR1 and SWIR2).

To manage the high atmospheric variability and occasional heavy cloud cover during the rainy season, a quality-control filter is applied to remove images with cloud cover that is higher than the selected biome threshold. This ensures a balance between atmospheric clarity and sufficient data density for the semi-arid region.

Unlike more temperate biomes, the Caatinga requires a compositing window that captures its extreme phenological contrast. Annual mosaics are generated using a standardized window typically spanning from specific peak-green to peak-dry intervals, ensuring the model can distinguish between the state of the dry season and the vigorous green state of the rainy season. This seasonal contrast is vital for reducing commission errors between native vegetation and pasture or agriculture.

Despite this advancement, we encountered a notable obstacle: the new mosaics exhibited a greater number of pixel gaps, particularly in areas of bare soil. To overcome this, we implemented a robust correction method. This involved utilizing mosaics from our previous methodology and applying linear regression on a band-by-band basis to approximate and align the two datasets. Consequently, we were able to seamlessly fill the pixel gaps in our new mosaics with accurate data derived from the earlier, more complete imagery (see example below):

Figure 3: Result of linear regression between the GEE mosaic and the MapBiomas mosaic.

The latest mosaic incorporates visible, infrared, and SWIR bands across the three aforementioned periods (annual, dry, and rainy seasons). Additionally, it includes descriptive statistics computed for the dry and wet periods, various spectral indices, and spectral mixing fractions, resulting in a comprehensive dataset of 142 bands.

3.1 DEFINITION OF THE PERIOD

To accurately classify LULC in the Caatinga biome, the project focused on selecting imagery that maximizes cloud-free coverage. This was critical due to the Caatinga's extreme phenological changes (seasonal leaf loss) driven by its unique, highly seasonal rainfall patterns**,** the period between January to July (with higher levels of rainfall). To define the best image selection periods for mosaic construction, they analyzed rainfall data for Brazil's Northeast region from INMET (1961-2015), accounting for the strong seasonality that directly impacts the vegetation's physiological activity. This dataset was obtained from the INMET (*www.inmet.gov.br*).

3.2 MOSAIC QUALITY

The final input for the Collection 3 classification in the Caatinga combines the Sentinel-2 Mosaics summarizing seasonal spectral behavior and the stark transition between the semi-arid wet and dry phases and annual satellite Embedding bands capturing contextual, structural, and spatial patterns within the landscape.

This integrated approach significantly enhances class discrimination in the structurally complex and highly seasonal Caatinga landscape. All final mosaics undergo rigorous visual inspection to ensure data suitability before being processed by the primary classifiers.

The base mosaic comprised six spectral bands from the Sentinel-2 sensor, specifically Blue, Green, Red, NIR, SWIR1, and SWIR2, across three distinct temporal periods: wet, dry, and annual. Each band was identified by a composite name, such as 'blue_median_dry', indicating the spectral band, the statistic (median), and the period (dry). Additionally, several spectral indices were calculated for the mosaics of each period and incorporated as additional bands.

Table 3.Additional bands added to the Caatinga feature space in the MapBiomas 10 m Collection 3 (beta).

Name	Formula	Description	Reference
RVI	(N * R) / (G ** 2.0)	Ratio Vegetation Index	Jordan (1969)
RATIO	N/R	Ratio	Pearson and Miller (1972)
NDWI	(G - N) / (G + N)	Normalized Difference Water Index	McFeeters (1996)
AWEI	B + 2.5 × G − 1.5 × (N + S1) − 0.25 × S2	Automated Water Extraction Index	Feyisa et al. (2014)
IIA	R/N	Inverse Intensity Index	
EVI	2.5 × (N + R) / (N + 6R - 7.5B + 1)	Enhanced Vegetation Index	Huete et al. (1994)
GCVI	(N/G) - 1	Green Chlorophyll Vegetation Index	Gitelson et al. (2005)
GEMI	(2 X (N^2 - R^2) + 1.5N + 0.5R)/(N + R + 0.5)	Global Environmental Monitoring Index	Pinty and Verstraete (1992)
CVI	(N x R) / (G^2)	Chlorophyll Vegetation Index	Vincini et al. (2008)
GLI	(2G - R - B)/(2G + R + B)	Green Leaf Index	Lourenço et al. (2021)
AVI	(N x (1- R) x (N - R))/ 1000	Advanced Vegetation Index	Loi et al. (2017)
BSI	((S1 + R) - (N + B))/((S1 + R) + (N + B))	Bare Soil Index	Rikimaru et al. (2002)
BRBA		Broadband Reflectance-Based Albedo	Liang (2001)
DSWI5	G / (N + S1 + R)	Dynamic Surface Water Index 5	Fisher et al. (2016)
LSWI	(N - S1)/(N + S1)	Land Surface Water Index	Xiao et al. (2002)
MBI	((S1 - S2 - N)/(S1 + S2 + N)) + 0.5	Modified Bare Soil Index	Nguyen et al. (2021)
UI	(S2 - N) / (S2 + N)	Urban Index	Kawamura et al. (1997)
OSAVI	(N - R) / (N + R + 0.16)	Optimized Soil Adjusted Vegetation Index	Rondeaux et al. (1996)

RI	(R - G)/(R + G)	Redness Index	Mathieu et al. (1998)
GVMI	((N + 0.1) - (S1 + 0.02)) / ((N + 0.1) + (S1 + 0.02))	Global Vegetation Moisture Index	Ceccato et al. (2002)
NIR_CONTRAST	1/14 GLCM metrics proposed by Haralick, Textural Features for Image Classification	CONTRAST NIR bands	Haralick et al. 1973
RED_CONTRAST	1/14 GLCM metrics proposed by Haralick, Textural Features for Image Classification	CONTRAST RED bands	Haralick et al. 1973
NDDI	(NDVI - NDWI) / (NDVI + NDWI)	Normalized Difference Drought Index	Gu et al. (2007)
NDVI	(N - R) / (N + R)	Normalized Difference Vegetation Index	Rouse et al. (1974)

3.3. DEFINITION OF REGIONS FOR CLASSIFICATION

Classifying homogenous regions helps reduce the spectral variability among pixels, both within and between LULC classes and allows the use of a consistent set of samples to classify large areas of the mosaic. However, it is a computationally expensive task. To address this, the Caatinga biome was divided into smaller areas based on watershed boundaries provided by the Agência Nacional de Águas (*www.ana.gov.br*) (Figure 4). The natural borders of the basins helped maintain the homogeneity of the areas and allowed for the automation of the sampling process using GEE's Python API. In earlier versions, level 4 watershed basins were selected, dividing the biome into 320 regions.

Due to changes in biome boundaries (IBGE, 2019), a merged version combining level 2 and level 4 watershed boundaries was employed, which reduced the Caatinga biome's division to 42 regions. In other words, watersheds that are already small at level 2 and were very fractionated at level 4 will remain with the level 2 polygon. In Collection 10.0, this division was further refined to 49 regions.

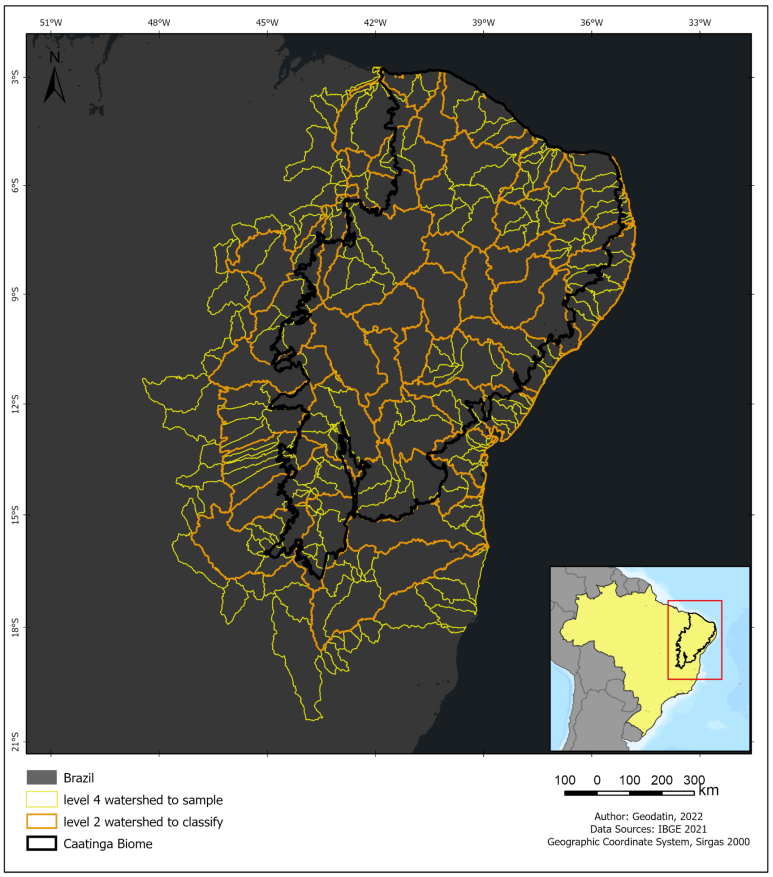

Figure 4. Watershed basins used in the classification and sampling of the MapBiomas LULC collections for Caatinga biome.

4. CLASSIFICATION PROCESS

4.1 SAMPLE SELECTION PROCESS

The most recent methodology of the sampling process aims to establish pixel collection areas with the least uncertainty in the label, for this purpose, exclusion and inclusion criteria for collection areas were established.

The exclusion criteria consider areas where there was some intra-annual change and could corrupt the annual spectral information. The changes considered are burned areas, deforested areas, areas within a buffer of gaps from clouds or cloud shadows and areas that show variability between consecutive years.

The inclusion criteria consider areas where as a likely sample pixel only in areas that were stable over a 3-year window. Another inclusion criterion was to consider those pixels with the same labels in 30-meters collection 10.0. To achieve these criteria for each region grid, sorting at least 500 samples per class was

required, which compelled the use of the function ee.Image().stratifiedSample() to collect samples from small areas inside a class.

The spectral information is essentially derived from the MapBiomas mosaic, but after analyzing the first set of samples, a significant number of other spectral indexes were calculated from the bands 'blue_median', 'green_median', 'red_median', 'nir_median', 'swir1_median', 'swir2_median' present in the mosaic. The new indexes calculated were the following: "ratio", "rvi", "awei", "iia", "gemi", “gvmi”,”gcvi”,”gsavi”, "cvi",”gli",”ndvi”,”ndti”,"afvi","avi","bsi","brba","dswi5","lswi","mbi","ui","osavi","ri","brightness", "wetness", "nir_contrast", "red_contrast".

In the 30-meter Collection 10.0, a new methodology was required due to the large volume of collected samples. The first step involves collecting samples from pixel areas with lower label uncertainty. The second step combines all samples within the 49 classification regions. The third step applies a downsampling method within each sample set. Collection 3 implemented the same sampling strategy to manage large-scale datasets.

Sample collection prioritizes areas where pixels are least likely to have misclassified labels. Several aforementioned criteria serve as filters for these collection areas. These areas were subsequently divided into 761 grids, covering the 49 basins (Figure 5). For Collection 3, 500 pixels were collected per grid for each class. This strategy resulted in an average of approximately 500,000 points per sample set (basin/year).

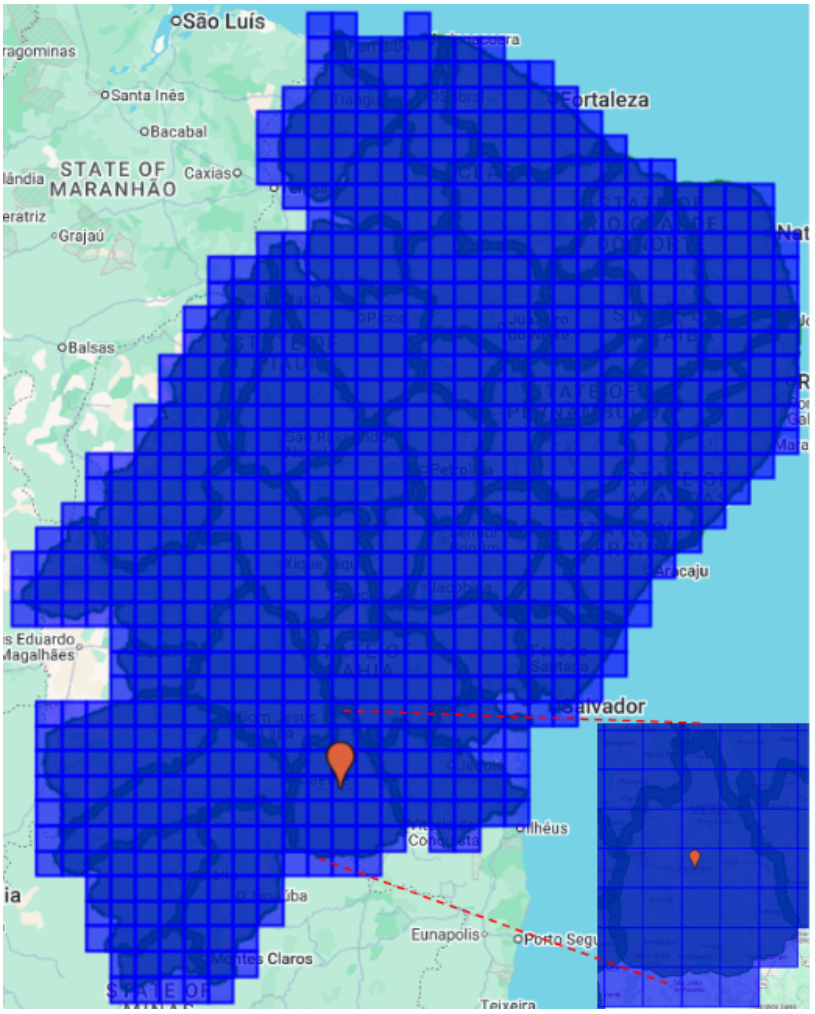

Figure 5. Collection areas by grid and their combined watershed limits.

The GTB classifier demands more computational power than RF, making it challenging to use very large sample volumes for training. This is precisely why we employ a downsampling process. This methodology allows us to extract a significantly smaller, yet representative, subset of the initial data, while also removing potential outlier pixels from the overall sample.

Inspired by the "Isolation Forest" algorithm (LIU et al., 2008), we developed a new methodology using a probabilistic variant of the GTB classifier". This approach enhances the selection of training samples, replacing the previous method that relied on the Google Earth Engine (GEE) API.

Our process begins by selecting subsets of samples for each year and for each basin within our dataset. The core idea is to refine the training data by identifying the most confident and representative pixels.

Here's how it works:

1. Initial Separation*:* We first separate samples belonging to the three primary natural vegetation classes:
    - Forest Formation
    - Savanna Formation
    - Grassland Formation
2. Probabilistic Output*:* For every pixel classified by the GTB model, the output isn't just a single class label. Instead, the model provides a *probability vector* indicating the likelihood of that pixel belonging to each class.

    Example: A probability vector like (0.2,0.85,0.12) means:
    - 20% chance of being Forest Formation
    - 85% chance of being Savanna Formation
    - 12% chance of being Grassland Formation
3. *Identifying High-Confidence Candidates:* Even though the pixel above would be ultimately labeled as "Savanna" (since 0.85 is the highest probability), the crucial insight comes from its **high confidence score** for that class. In the *feature space* (the multi-dimensional representation of the pixel's characteristics), the classifier effectively "sees" this pixel as having an 85% probability of being Savanna. This makes it an ideal candidate for inclusion in the training set for the Savanna class.

To significantly optimize and reduce the size of the overall training dataset, we applied a strategic selection process:

- For each of the natural vegetation classes, we specifically chose 100 pixels whose classification probabilities fell within high-confidence intervals: (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0). This ensures that only the most confidently classified pixels are retained for training.

This same refined approach was also extended to agricultural classes, including:

- Pasture
- Agriculture
- Mosaic of Uses (mixed land use)

Classes with low representation in the initial dataset, such as Water, Other Non-Vegetated Areas and Rocky Outcrop, were not included in this specific sampling strategy. They had too few samples to meaningfully apply this high-confidence selection method.

4.4 FEATURE SPACE AND FEATURE SELECTION PROCESS

The feature space for digital classification of the LULC classes in the Caatinga biome comprised a subset of 75 features (Figure 6), taken from the complete feature space of MapBiomas Collection 7.0 (General ATBD MapBiomas, 2020). In Collection 3, a larger number of spectral indices were calculated to expand the feature space of the MapBiomas mosaic. The goal was to find a reduced space that offers more separability and contrast between targets.

Bands	Estimators	Index Spectral	Estimators	Franctions	Estimators
blue	median	CAI	median	gv	amp
	median dry		median dry		median
	median wet		stdDev		media dry
	min	EVI2	amp	npv	median
green	median		median		median dry
	median dry		media dry		median wet
	median wet		stdDev		min
	median texture	GCVI	median	soil	median
	stdDev		median dry		median dry
red	median		median wet		median wet
	median dry	NDVI	amp		stdDev
	median wet		median	ndfi	median
	min		median dry		median dry
nir	median		median wet		median wet
	median dry	NDWI	amp		min
	median wet		median	sefi	median dry
	min		median dry		median wet
SWIR1	median		median wet		stdDev
	median wet	SAVI	median	shade	median
	min		median dry		median dry
	stdDev		median wet		median wet
SWIR1	median		stdDev		min
	median wet	PRI	median		amp
	min		median dry		
	stdDev		median wet		

Figure 6: Feature space subset considered in the classification of Landsat image mosaics in the Caatinga biome in the MapBiomas Collection 3.

The feature space of this collection has been expanded to be more robust and to follow good data augmentation practices used in data science (Figure 7).

Index Spectral	Estimators	Index Spectral	Estimators	Index Spectral	Estimators
RATIO	median	GLI	median	LSWI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
RVI	median	AFVI	median	MBI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
GEMI	median	AVI	median	UI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
AWEI	median	BSI	median	OSAVI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
IIA	median	BRBA	median	RI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
CVI	median	DSWI5	median	Brightness	median
	median dry		median dry		median dry
	median wet		median wet		median wet
GVMI	median	NIR Contrast	median	Wetness	median
	median dry		median dry		median dry
	median wet		median wet		median wet
Red Contrast	median				
	median dry				
	median wet				

Figure 7: Feature space subset indexes calculated from the estimated bands of the Landsat mosaic of mapBiomas in the Caatinga biome in the MapBiomas Collection 10.

The image below (Figure 8) depicts an instance of the samples corresponding to sub-basin “744” which have an unbalanced distribution due to the nature of the data.

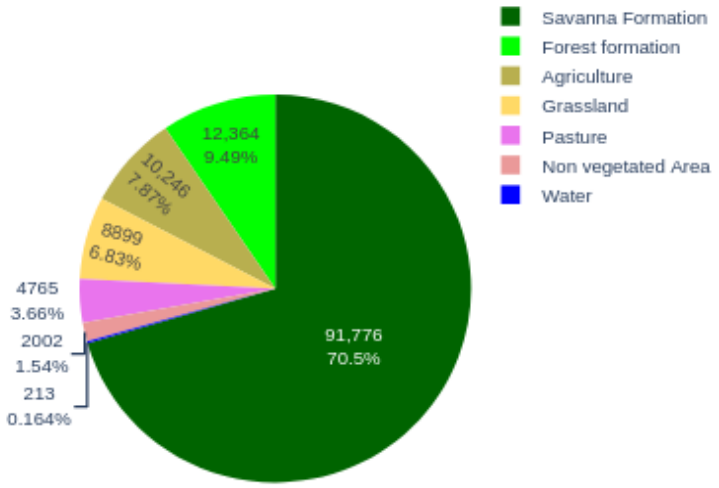

Figure 8: Distribution of samples for sub-basin 744 in the year 2000.

Achieving separability in the feature space is a prevalent challenge when performing remote sensing image classification in the Caatinga Biome. Figure 9 demonstrates that separability within a spectral band is limited for various targets in the image. Another way of visualizing this can be seen in Figure 10, which plots the "blue_median", "green_median", "red_median", "nir_median" bands of the mosaic for six coverage classes.

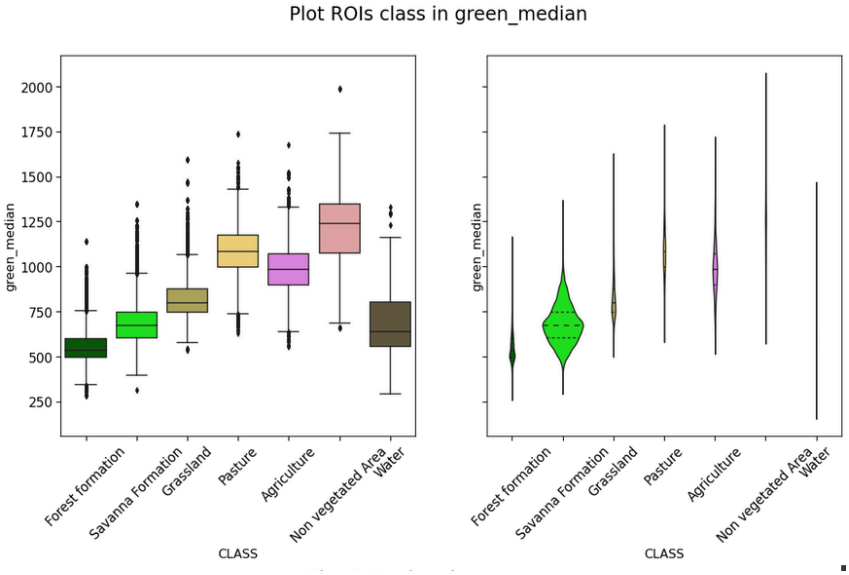

Figure 9: Box and violin plots from samples of spectral band “GREEN” in the main land cover classes mapped by the Caatinga team.

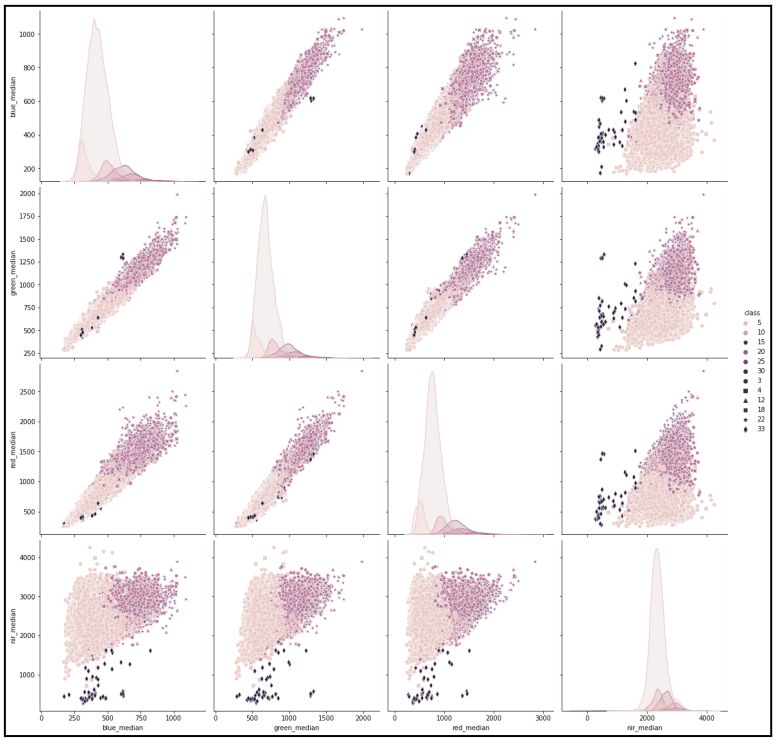

Figure 10: Spatial distribution of samples for the variables “blue_median”, “green_median”, “red_median”, “nir_median”.

All watersheds were analyzed individually in terms of feature importance. These variables included the original Landsat reflectance bands, as well as vegetation indexes and spectral mixture modeling-derived variables. The first step was measuring the correlation between feature Collection variables (Figure 11), and some variables would be eliminated from the least important criteria following the score.

To calculate the correlation between the indices, the *corr()* function was used for each set of samples. The *corr()* function is implemented in the Pandas library of the python language. The python scripts were implemented in colab.

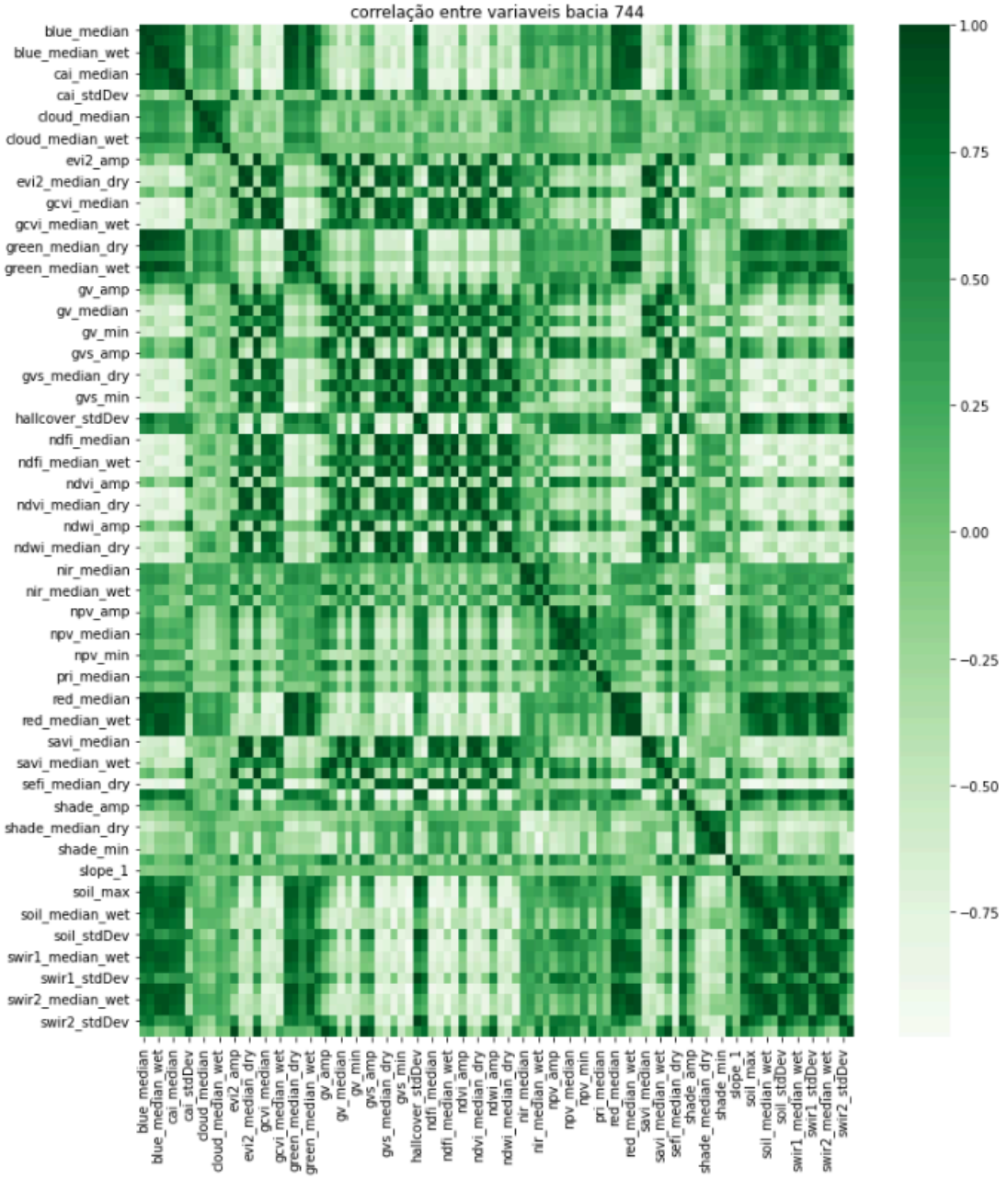

Figure 11. Example the plot correlation of watersheds samples from the year 2020.

As with the previous Collections, the model has included Recursive Feature Elimination with Cross Validation (RFECV), an alternate feature selection method that uses cross-validation to automatically optimize the amount of features picked. As a result, for each set of data (basin / year), a list of characteristics chosen during the feature removal procedure was saved (*ZHANG AND JIANWEN, 2009; RAMEZAN, 2022*). A basic example may be found at the link below:

https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html

The *RFECV()* function can be accessed by using the python Sklearn library (Figure 12). There are two methods in which the class can be used to filter the selected variables: the "support_()" method and the "ranking_" method. With the former we can choose the surviving variables from a list of "TRUE" or "FALSE", and with the latter we can extract the ranking of the "TRUE" variables.

If the number of variables in "TRUE" is less than 10, then banking consecutive to 1 is taken as a condition (e.g. 2,3,4,5 etc.).

```
def method_RFECV(self, X_train, y_train, nameExports):
    # namebacia = nnameFile.split('_')[0]
    # myear = nnameFile.split('_')[1]
    skf = RepeatedStratifiedKFold(n_splits=12, n_repeats=5, random_state=36)
    model = GradientBoostingClassifier()
    min_features_to_select = 6
    rfecv = RFECV(
            estimator=model,
            step=1,
            cv= skf,
            scoring= 'accuracy',
            min_features_to_select=min_features_to_select,
            n_jobs= 8
        )

    rfecv.fit(X_train, y_train)
    dict_inf = {
        'features': X_train.columns,
        'rankin': rfe.ranking_,
        'support': rfe.support_
    }

    rf_df = pd.DataFrame.from_dict(dict_inf)
    namePathtmp = self.namepathroot + '/' + self.nameFolderSaved+ '/' + 'rfeCVOut_' + nameExports
    rf_df.to_csv(namePathtmp, index=False, sep=';')
```

Figure 12: Example of the implemented feature selection function (RFECV) and a list of selected variables.

4.5 HYPERPARAMETER TUNING PROCESS

A script was implemented for the Hyperparameter Tuning process after selecting the variable sets by drainage basin and year. The GridSearchCV() function, along with the Pipeline() function, is capable of testing various parameter combinations for the model. It is then possible to establish which combination of parameters represents the best score or accuracy. The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid. An example of the "learning rate" parameters and "n estimators" is shown in figure 13, where the optimal pair of parameters would be (40, 0.175).

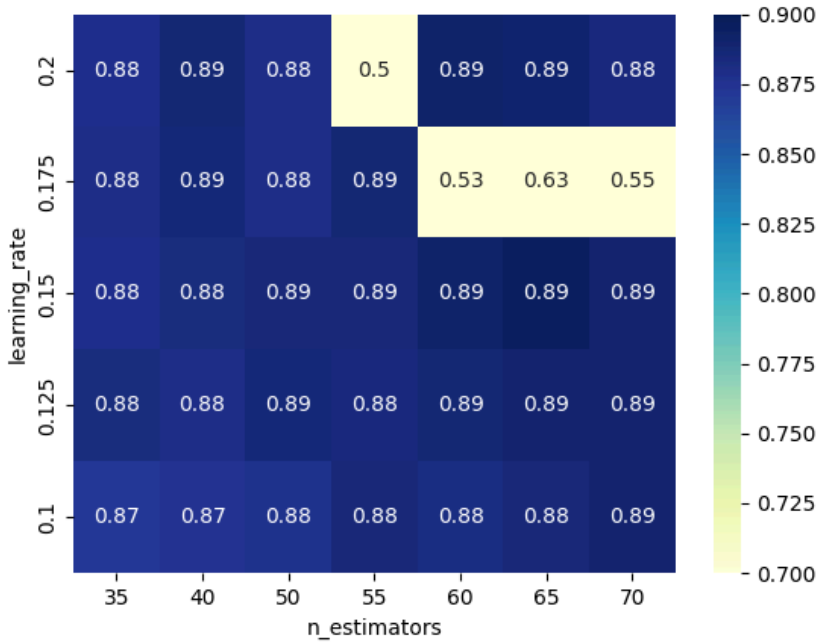

Figure 13. Example of the plot of combination of "learning rate" parameters and "n estimators".

The GridSearchCV() (Grid Search Cross-Validation) is an exhaustive, or "brute-force," hyperparameter optimization technique, which means it can be computationally expensive. It operates as follows:

1. Defining a Hyperparameter "Grid": You define a dictionary where keys are the hyperparameter names for your model, and values are lists of all possible values you want to test for each hyperparameter. This creates a "grid" of every possible combination. (See Figure 14 for an illustration).
2. Exhaustive Training and Cross-Validation: For each hyperparameter combination within your defined grid, `GridSearchCV()` performs the following steps:

- It trains the model using cross-validation. This involves splitting the training dataset into k "folds" (subsets). The model is then trained k times; each time, it uses k-1 folds for training and one fold for validation.
- The model's performance is evaluated on each validation fold using a pre-defined scoring metric (e.g., accuracy, F1-score, Mean Squared Error (MSE), etc.).
- The final score for that specific hyperparameter combination is the average of the scores obtained across all k folds.

3. Selecting the Best Combination: After evaluating all possible combinations in the grid via cross-validation, `GridSearchCV()` identifies the combination of hyperparameters that yielded the best average validation score.
4. Final Model Refitting: Once the optimal hyperparameter combination is found, `GridSearchCV()` (by default, if `refit=True`) retrains the model using the entire original training dataset with these winning hyperparameters. This ensures you have a robust final model trained on all available data.

Part of the code implemented for selecting optimal parameters is shown in the following image (Figure 14). Each pair of optimal parameters for year and hydrographic region is saved in a single json file.

```
# random_state=0,
model = Pipeline([
            ("classifier", ensemble.GradientBoostingClassifier(
                                n_estimators= 150,
                                learning_rate= 0.01,
                                subsample= 0.8,
                                min_samples_leaf= 3,
                                validation_fraction= 0.2,
                                min_samples_split= 30,
                                max_features= "sqrt"
                            ))
        ])
print("Modelo Pipeline ", model)

param_grid = {
    'classifier__learning_rate': (0.1, 0.125, 0.15, 0.175, 0.2),
    'classifier__n_estimators': (35,40, 50, 55, 60, 65, 70)
}
model_grid_search = GridSearchCV(
                            model,
                            param_grid=param_grid,
                            n_jobs=2,
                            cv=2
                        )
model_grid_search.fit(data_train, target_train)

accuracy = model_grid_search.score(data_test, target_test)
print(
    f"The test accuracy score of the grid-searched pipeline is: {accuracy:.2f}")

model_grid_search.predict(data_test)

print(f"The best set of parameters is: "
    f"{model_grid_search.best_params_}")
```

Figure 14: Part of the code implemented for the Hyperparameter tuning process.

For each watershed sample, a list of variables was kept for eventual use in the classification process. All the codes used in this stage are available in the repository of MapBiomas's Github (https://github.com/mapbiomas-brazil/caatinga).

4.6 CLASSIFICATION ALGORITHM

During the classification process, the input data is adjusted to allow the MapBiomas mosaics to be classified by hydrographic basin and year. The data is then displayed using a GEE script and reviewed by the team's analysts to assess the classification results by basin and year. The primary objective of this step is to identify regions that require additional samples or classification parameter changes. Once identified, these areas are included in the map correction cycle as a result of the GTB classification (LAWRENCE et al. 2004). An example of the parameters for GTB classifiers is shown in figure 15.

```
# https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting
'pmtGTB': {
    'numberOfTrees': 45,
    'shrinkage': 0.1,
    'samplingRate': 0.8,
    'loss': "LeastSquares",#'Huber',#'LeastAbsoluteDeviation',
    'seed': 0
},
```

Figure 15: Example parameters for the Gradient Tree Boost classifiers.

For the classes of *Forest Formation, Savanna Formation, Grassland Formation, Pasture, Agriculture, Mosaic of Uses, Other Non-Vegetated Areas, and Water Bodies,* the GTB classifier is applied. This process uses a specific sample set, a predefined list of spectral bands, a set of classifier parameters, and a dictionary indicating the percentage of samples per class for each region/year processed in the construction of the map series that compose the collection.

Each classification version undergoes a set of stringent criteria, including accuracy, smoothness of area curves across the entire series, and spatial coherence. These criteria determine whether the map series for a given region demonstrates superior quality compared to the same region in previous collections.

The *Rocky Outcrop* class presents significant mapping challenges due to its inherent spectral mixture, often combining exposed soil with sparse grasses or small plants growing amidst the rocks. To address this complexity, our mapping effort

leveraged data from the Geological Survey of Brazil (SGB), which provides presence points for these outcrops.

We demarcated these outcrop areas using polygons that encompassed over 80% of the outcrop's extent. Within these polygons, we applied an unsupervised clustering algorithm, specifically `ee.Clusterer.wekaXMeans` implemented in Google Earth Engine (GEE), configured to produce a maximum of three clusters, PELLEG AND MOORE (2000). From the resulting raster output, we identified the cluster value corresponding to the rocky outcrop class and then selected all such areas across the Caatinga Biome.

While this clustering process for identifying rocky outcrops is time-consuming, it proved fundamental. It provided a robust foundation for constructing a high-quality dataset of labeled image patches (or "chips") essential for training subsequent Deep Learning models.

4.6.1 GRADIENT TREE BOOSTING

Gradient Tree Boost (GTB) is a boosting ensemble method that builds decision trees sequentially. Each new tree built in sequence is used to correct the errors (residuals) of the preceding tree, with the goal of minimizing a specific loss function.

Advantages of Gradient Tree Boosting:

- *High Accuracy:* Frequently achieves state-of-the-art accuracy in many problems, outperforming Random Forest in some cases, especially when well-tuned.
- *Ability to Capture Complex Patterns:* By iteratively correcting errors, it's very good at capturing non-linear relationships and complex interactions in the data.
- Handles Imbalanced Data Well: Can handle imbalanced datasets more effectively by focusing on difficult-to-predict examples (those with larger residuals). This property fits perfectly into the challenge of classifying 8 cover classes within a large volume of Landsat images.
- Flexibility: Can be optimized for a variety of loss functions, making it applicable to various problem types (regression, classification, ranking, etc.).

- Feature Importance: Similar to Random Forest, it also provides measures of feature importance. This property is widely used in the selection of bands and spectral indices to be used during the final classification process.

Disadvantages of Gradient Tree Boosting:

- *More Prone to Overfitting*: Because it builds trees sequentially and focuses on errors, GTB is more susceptible to overfitting, especially on noisy data or if hyperparameters aren't tuned carefully. This property makes the process of acquiring training samples, the feature selection process and the hyperparameter tuning process rigorous in this work.
- Slower to Train: Training is sequential, meaning it cannot be parallelized as easily as Random Forest, resulting in longer training times. Therefore, the construction of many trees within the GTB causes the processing on the GEE platform to time out memory.
- Sensitive to Hyperparameters: Requires more careful and extensive tuning of hyperparameters (such as learning rate, number of trees, and maximum depth) to achieve optimal performance.
- Sensitive to Outliers: By focusing on reducing residuals, outliers can have a disproportionate impact, leading the model to "learn" the noise.

We prioritized using Gradient Tree Boosting because:

- Maximum accuracy is crucial due to working with highly seasonal data, specifically annual mosaics for the Caatinga biome. Therefore, even though it requires more time to fine-tune the hyperparameters, we selected the model that achieves the best accuracy.
- New methodologies for cleaning the training datasets ensure that the model performs better than previous models.
- Imbalanced datasets are a natural property of remote sensing data, and GTB can be more effective at learning minority classes by focusing on errors.

5. POST-CLASSIFICATION

The mosaics used in the classification process exhibit noise and pixel gaps, which are more frequent at the beginning of the series with Landsat 5 data. These gaps, predominantly in the early years of the series, are corrected through temporal, frequency, and spatial filters, applied in the following order:

● Gap-Fill Filter: This operation fills pixels with invalid values, using valid pixels from adjacent times (t+1 or t-1), following the chronological sequence of years (ascending or descending).

● Frequency Filter: Applied to natural classes, this filter stabilizes small seasonal changes, especially in savannas and grasslands, which historically show greater confusion in the collections. It works by applying rules that stabilize the time series: pixels classified as savanna in 85% of the series force adjacent pixels of Forest Formation or Grassland Formation to be reclassified as savanna. Similar rules are applied for pixels with a frequency of 90% forest and 80% Grassland Formation.

● Temporal Filter: This filter corrects short-term inconsistencies (1-2 years) in the time series. Pixels with atypical coverages, relative to the environmental pattern, are corrected by a moving window of 3, 4, or 5 pixels.

● Spatial Filter: This filter eliminates "salt and pepper" noise, which are isolated pixels in areas of homogeneous classes. The operation ensures spatial coherence, using the "connectedPixelCount" function of Google Earth Engine (GEE) to assign a pixel the predominant class among its neighbors. After each cycle of applying these filters, one or more steps may be repeated in the post-classification, to refine the correction."

The temporal filter rules were specifically adapted to account for the phenological and spectral dynamics of the land cover classes characteristic of the Caatinga biome. In addition to the standard temporal consistency checks, custom rules were incorporated to handle anomalous transitions, particularly cases in which a class appears abruptly in a pixel’s time series. These adjustments aim to improve classification stability and reduce spurious temporal fluctuations in areas with high seasonal variability and low vegetation cover.

6. INTEGRATION

The final classification results, incorporating post-classification filters, are integrated with the data from the cross-cutting themes across the entire historical series (1985–2024). This step is carried out based on predefined prevalence rules (Table 3). Ultimately, the final integrated map for the Caatinga biome features 20 classes at level 3 of the Collection 10.0 legend (Figure 16).

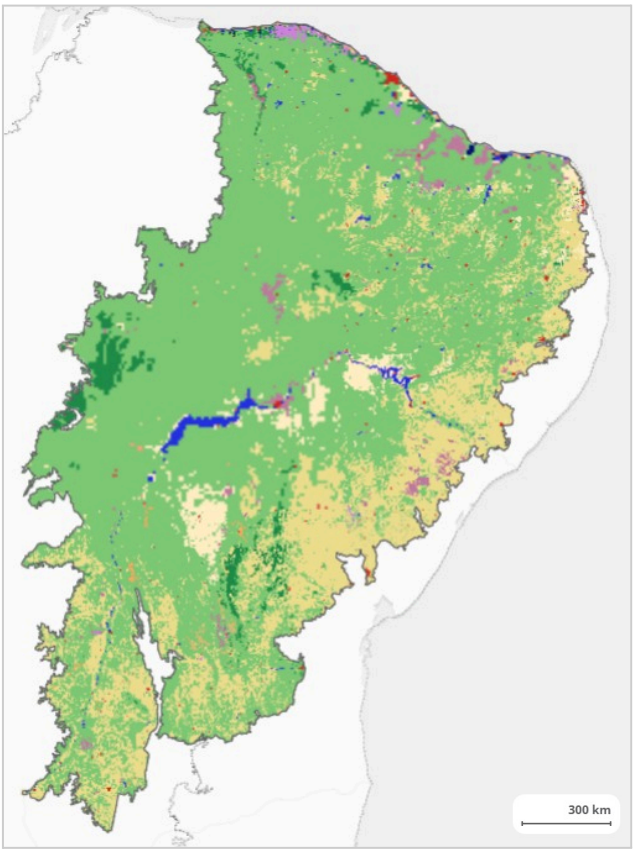

Figure 16. Final land use and land cover map of the Caatinga biome (2024).

Table 4. General prevalence rules - MapBiomas Collection 3

Class	ID	Prevalence order	Color
Photovoltaic power plant	75	1	
Mining	30	2	
Beach, Dune and Sand Spot	23	3	
Mangrove	5	4	
Aquaculture	31	5	
Hypersaline Tidal Flat	32	6	
Urban Infrastructure	24	7	

Class	ID	Prevalence order	Color
Forest Plantation	9	8	
Rocky Outcrop	29	9	
Temporary Crops	19	10	
Perennial Crops	36	11	
Herbaceous Sandbank Vegetation	50	12	
River, Lake and Ocean	33	13	
Other non Vegetated Areas	25	14	
Forest Formation	3	15	
Savanna Formation	4	16	
Wooded Sandbank Vegetation	49	17	
Wetland	11	18	
Grassland Formation	12	19	
Pasture	15	20	
Mosaic of Uses	21	21	

8. REFERENCES

ARCOVA, F. C. S.; CICCO, V. DE; ROCHA, P. A. B. Precipitação efetiva e interceptação das chuvas por floresta de Mata Atlântica em uma microbacia experimental em Cunha - São Paulo. Revista Árvore, v. 27, n. 2, p. 257–262, 2003.

Breiman, 2001. Classification and regression based on a forest of trees using random inputs. doi:10.1023/A:1010933404324.

IBGE. Vegetação RADAM. Disponível em: <ftp://geoftp.ibge.gov.br/informacoes_ambientais/acervo_radambrasil/vetores/>. Acesso em: 30 maio. 2018.

IBGE Mapa de Biomas do Brasil – primeira aproximação. Rio de Janeiro, 2020, disponível em: https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html?=&t=download s, acessado em julho de 2020;

LAWRENCE, R., BUNN, A., POWELL, S., & ZAMBON, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote sensing of environment*, *90*(3), 331-336.

LIU, F.T., TING, K.M. AND ZHOU, Z.H., 2008, December. Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

PATIL, D., PATIL, K., NALE, R. AND CHAUDHARI, S., 2022, July. Semantic segmentation of satellite images using modified U-Net. In *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-6). IEEE.

PELLEG, D. AND MOORE, A.W., 2000, June. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml* (Vol. 1, pp. 727-734).

TORTORA, R.D. 'A Note on Sample Size Estimation for Multinomial Populations." The American Statistician 32:3 (August 1978), 100-102.

T. KOHONEN, "Learning Vector Quantization", The Handbook of Brain Theory and Neural Networks, 2nd Edition, MIT Press, 2003, pp. 631-634.

ZHANG, RUI, AND JIANWEN MA. "Feature selection for hyperspectral data based on recursive support vector machines." International Journal of Remote Sensing 30.14 (2009): 3669-3677.

RAMEZAN, CHRISTOPHER A. "Transferability of Recursive Feature Elimination (RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification." Remote Sensing 14.24 (2022): 6218.

RONNEBERGER, O., FISCHER, P., & BROX, T. 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

ZHU, X.X., TUIA, D., MOU, L., XIA, G.S., ZHANG, L., XU, F. AND FRAUNDORFER, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, *5*(4), pp.8-36.