Revista Contexto Geográfico

ISSN: 2595-7236

Artigo | Article

Machine learning on alluvial deposit Identification in dryland area

Aprendizado de máquina na identificação de depósitos aluvionares em áreas de clima árido

Daniel Vio
Mestre em Ciência da Computação
Universidade Federal da Paraíba – UFPB, João Pessoa, Brasil
e-mail, daniel.vio@academico.ufpb.br

Jonas Otaviano Praça de Souza
Doutor em Geografia
Universidade Federal da Paraíba – UFPB, João Pessoa, Brasil
e-mail, jonas.souza@academico.ufpb.br

Gustavo Henrique Matos Bezerra Motta
Doutor em Engenharia Elétrica
Universidade Federal da Paraíba – UFPB, João Pessoa, Brasil
e-mail, gustavo@ci.ufpb.br

Leandro Carlos de Souza
Doutor em Ciências da Computação
Universidade Federal da Paraíba – UFPB, João Pessoa, Brasil
e-mail, leandro@ci.ufpb.br

Abstract
There is a growing global concern about water resources and an increasing interest in studies identifying groundwater. Alluvial aquifers exhibit varied forms and irregular distribution in the landscape, making their location challenging. This study applies machine learning (ML) techniques to detect alluvial deposits in the Riacho do Tigre watershed in the semi-arid region of northeastern Brazil. Fourteen input variables and one output variable were collected across approximately one and a half million points distributed along the main channels of the watershed. Using decision trees (DT), the model was trained and validated through k-fold cross-validation and bootstrap methods, achieving an accuracy of 0.92, indicating good performance in classifying alluvial areas.

https://doi.org/10.28998/contegeo.10i.24.18434

Submetido em: 06/11/2024
Aceito em: 31/03/2025
Publicado: 09/09/2025

e-Location: 18434

***Keywords**: Hydrology; Artificial intelligence; Decision tree; Alluvial areas, Remote sensing.*

Resumo

Há uma crescente preocupação global com os recursos hídricos e um interesse cada vez maior em estudos voltados para a identificação de águas subterrâneas. Os aquíferos aluviais apresentam formas variadas e distribuição irregular na paisagem, o que dificulta sua localização. Para enfrentar essa dificuldade, o presente estudo aplica técnica de aprendizado de máquina para detectar depósitos aluviais na bacia hidrográfica do Riacho do Tigre, localizada no semiárido nordestino brasileiro. Foram coletadas catorze variáveis de entrada e uma variável de saída em aproximadamente um milhão e meio de pontos distribuídos pelos canais principais da bacia. Utilizando árvores de decisão (DT), o modelo foi treinado e validado através de validação cruzada k-fold e bootstrap, obtendo uma acurácia de 0,916, indicando um bom desempenho na tarefa de classificar áreas de aluvião.

Palavras-chave: Hidrologia; Inteligência artificial; Árvore de decisão; Áreas aluvionares; Sensoriamento remoto.

INTRODUÇÃO

The water deficit in dryland regions directly affects surface water availability, including spatial and temporal continuity of river flow (SOUZA, 2015). In this configuration, surface water flow occurs predominantly during precipitation events, characterised as ephemeral rivers or seasonality during the rainy season, with the appearance of intermittent rivers (MCLEOD et al., 2024). At the same time, droughts historically affect socio-economic development in these areas. In those scenarios, even surface water reserves, such as lakes and artificial reservoirs, could be depleted (BÚRQUEZ et al., 2024). In contrast, underground water reserves, including alluvial aquifers, are fundamental resources during dry/drought periods (SILVA and SOUZA, 2023).

Alluvial aquifers are formed through the erosion process of slopes, which carry various sediments (sand, silts, clay) to the beds and banks of channels. These sediments are loci of alluvial water deposits with high infiltration capacity and protection against evaporative effects. At the same time, the low-cost techniques to explore these shallow aquifers emphasise their importance to isolated rural communities (RITCHIE et al., 2021), especially where there is a predominant crystalline rock basement and fissure aquifers. The fissure aquifers generally have low hydrological potential, small volume and low water quality (SILVA and SOUZA, 2023). Due to these characteristics, alluvial aquifers are a strategic element for water supply and socio-economic development in those dryland areas.

There is a significant increase in global concern about threats to water resources (TAYER et al., 2023), including alluvial aquifers. The research conducted by Jasechko et al. (2024) involved measurements in thousands of wells and aquifers in various arid and semi-arid regions. The authors point out a global trend where many aquifers have experienced a rapid decline in recent years, primarily driven by excessive groundwater withdrawals, particularly for irrigation in arid and semi-arid regions, as well as reduced recharge rates due to decreasing precipitation and climate variability. In this context, there is a growing awareness of the importance of increasing knowledge about water reserves in dry areas.

Given the complexity of identifying and monitoring these resources, Machine Learning (ML) models are increasingly applied to identify groundwater, including alluvium. Indeed, artificial intelligence techniques offer significant potential for modelling complex systems in studies related to water resources and other fields. This approach eliminates the need to establish mathematical relationships between variables or physical parameters of the system, as it can build these relationships during the training, testing, and validation processes (ZARESEFAT and DERAKHSHANI, 2023).

The identification and mapping of alluvial aquifers present a significant challenge due to their dispersion across the landscape and the absence of a clearly defined geometric shape (CERVI and TAZIOLI, 2021). This complexity further complicates the task. However, ML techniques prove promising as they allow for identifying patterns in the data, enabling the efficient and accurate mapping of alluvial deposits.

This research focuses on the Riacho do Tigre watershed, located in the semi-arid region of Northeast Brazil. This study aims to detect the presence of alluvial areas along the three main rivers of the basin. Approximately one and a half million points were distributed along the channels, with each point assigned fourteen input variables and one output variable, characterising the area as alluvial. Following data collection and preprocessing, the research employed a ML technique, specifically decision trees (DT) (QUINLAN, 1986), to obtain the results.

DECISION TREES

The advancement in computational power in recent years has increased the relevance of ML. One of the main challenges algorithms face in this field is maximising their generalisation capacity, which means providing efficient responses

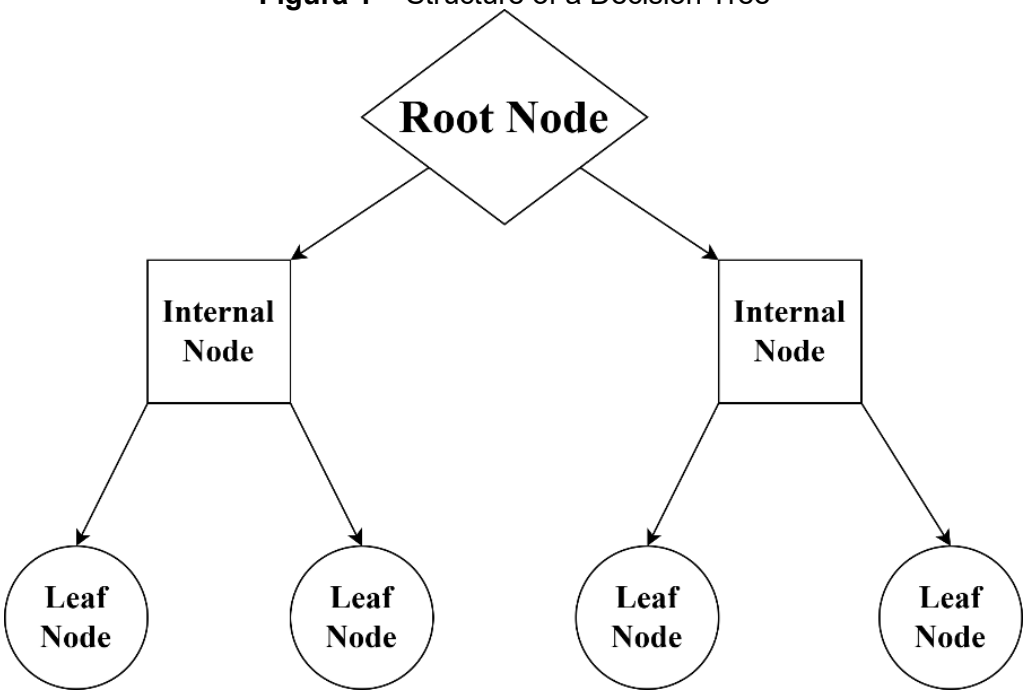
in situations not encountered during the learning process (LIMA, 2014). This capacity is crucial to ensure systems make appropriate decisions in different scenarios (MITCHELL, 1997).

The ML field can be categorised into four fundamental paradigms: Supervised Learning, Reinforcement Learning, Unsupervised Learning, and Semi-Supervised Learning. This research focuses on the development of a supervised DT. In this paradigm, algorithms are trained with a labelled dataset to learn a function that accurately maps inputs to their corresponding outputs (RUSSEL, 2022).

DT helps represent knowledge acquired from datasets, organising themselves as a combination of constraints on the attribute values of instances. Each path from the root to a leaf of the tree represents a sequence of tests on attributes, where the internal nodes represent decision points, and the branches indicate the different possible outcomes.

During the construction of the DT (Figure 1), the selection of attributes for splitting is determined by the purity of each node. It begins with the root node to create a more compact and efficient classification tree (MITCHELL, 1997). This process involves the recursive selection of attributes to form internal nodes and branches corresponding to each possible value of the selected attribute.

Figura 1 – Structure of a Decision Tree



Drafting: Authors (2024)

The construction continues until the examples in a node are homogeneous, creating leaf nodes, where the final classification is made. The decision regarding which attribute to split is based on the choice that results in a smaller and more accurate DT, as measured by the purity of each node along the path.

This research utilised the CART (Classification and Regression Trees) algorithm, which applies to classification and regression problems. A fundamental characteristic of this algorithm in classification is the use of the information metric known as the Gini index. The Gini index is calculated as:

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

Where n corresponds to the number of classes, pi represents the proportion of class i at the node. The closer the Gini index is to 0, the greater the purity of the node; as it approaches 0.5, the impurity increases.

The Gini index ensures that each node represents a set of instances belonging to the same class, thus avoiding inconsistencies. When deciding which feature to use for splitting the data, the algorithm examines the distribution of instances among the classes to determine the purity of a node and select the best feature for the split (MARSLAND, 2015).

Resampling methods are crucial tools for validating DT models. In this research k-fold cross-validation (HASTIE et al., 2009) and Bootstrap (EFRON and TIBSHIRANI, 1993) were employed. In k-fold cross-validation, the observations are divided into k groups or folds. In each iteration, one group is retained as validation data to test the model, while the remaining groups are used for training. This process is repeated k times, ensuring each group is used once as validation data. In the end, the mean error of all iterations is calculated, providing a more accurate estimate of the models. Bootstrap, on the other hand, involves creating multiple random samples with replacements from the original dataset, allowing for the assessment of the model's variability and robustness

MACHINE LEARNING AND DETECTION OF ALLUVIUMS

Identifying and mapping alluvial areas play a significant role in water resource management. The formation of alluvial aquifers occurs through the fragmentation of rocks and the transportation of sediments by rainfall. These sedimentary deposits are essential for water retention in arid regions, acting as natural reservoirs that can ensure water sustainability during periods of drought (BRAGA et al., 2016).

Historically, the identification of alluviums depended on direct observations and geological field mapping. Initially, research was based on visual descriptions and manual analyses of soil samples to understand the composition and extent of alluviums, approaches that were time-consuming and costly as they included inspections and topographic surveys (AMIT et al., 1996).

The transition to more advanced methods began with remote sensing technologies. Using satellite imagery and synthetic aperture radar (SAR) transformed the ability to map and analyse large alluvial areas more precisely. Farr and Chadwick (1996) demonstrated the effectiveness of these technologies by using SAR data to map alluvial fans in the Kun Lun Mountains, China, enabling the mapping of the morphology of alluvial fans and the identification of geomorphic processes. Similarly, Hertz et al. (2016) and Gaber et al. (2010) utilised SAR to analyse alluvial surfaces in deserts. Zhang et al. (2013) combined SAR and Digital Elevation Models (DEMs) to map alluvial fans.

Goorabi et al. (2021) and Iacobucci et al. (2024) utilise DEMs to enhance the geomorphological mapping of alluvial fans in arid regions, applying quantitative methods such as morphometric analysis and topographic feature extraction to identify and characterise these formations.

Crouvi et al. (2006) used field spectrometry and hyperspectral remote sensing in the Negev Desert, Israel, to identify specific spectral signatures associated with developing desert pavements and accumulating rock coatings, with an accuracy margin of approximately 15%. The introduction of multispectral techniques significantly impacted the identification of alluviums. Gillespie et al. (1984) used NASA's Thermal Infrared Multispectral Scanner (TIMS) to map alluvial fans in Death Valley, California, demonstrating the effectiveness of thermal sensing in discriminating between different

sediments. Al-Juaidi et al. (2003) investigated the fusion of remote sensing data to map geomorphological features and date alluvial surfaces in Saudi Arabia.

The introduction of Light Detection and Ranging (LiDAR) has added a new dimension to the identification of alluviums, enabling the creation of accurate DEMs. These models facilitate detailed analysis of the topography and structure of alluviums, allowing for the identification of geomorphological features with unprecedented precision. Hohenthal et al. (2011) and Cavalli et al. (2008) highlight that LiDAR has become an efficient tool for obtaining detailed topographic information, even in mountainous and densely forested areas.

Pioneering studies by Staley et al. (2005) and Frankel and Dolan (2007) demonstrated the effectiveness of LiDAR in analysing deposition patterns and characterising the roughness of alluvial fan surfaces. These studies revealed distinct deposition zones and allowed for differentiation between alluvial fan surfaces of varying ages. Subsequent research, such as that by Cavalli and Marchi (2008) and Regmi (2014), also employed LiDAR in identifying and classifying alluvial fans.

The convergence of such studies and machine learning (ML) methods has gained prominence in recent years in hydrological research. As Muñoz-Carpena et al. (2023) highlighted, integrating these approaches can reduce the uncertainty of hydrological models and improve the accuracy of predictions, particularly in large integrated systems.

Two types of studies on groundwater using machine learning techniques stand out in the literature. The first focuses on groundwater detection, with examples including the works of Díaz-Alcaide and Santos (2019), Ali et al. (2023), Seifu et al. (2023), Martinez-Santos and Renard (2020), and Nguyen et al. (2020). The second type involves developing models specifically for groundwater level prediction, such as the studies by Ardabili et al. (2020), Tao et al. (2022), UC-Castillo et al. (2023), Gholami et al. (2023), Kayhomayoon et al. (2022), Vadiati et al. (2022), Shakya et al. (2022), Srivastava et al. (2023), Gaffor et al. (2022), Luiz (2022), and El Bilali et al. (2021).

While groundwater research has attracted substantial attention, applying ML techniques to identify and classify alluvial formations, such as alluvial fans, has received less focus. A few researchers have explored this area, but these few studies have contributed valuable insights into the delineation and characterisation of alluvial areas.

Pipaud and Lehmkuhl (2017) conducted a study that presents a method for delineating and classifying alluvial fans using DEMs combined with the mean-shift clustering technique and a support vector machine (SVM). The input variables used in the segmentation included morphometric parameters such as slope, transverse curvature, longitudinal curvature, asymmetry of altitude values, and the gradient deviation from the fan apex. The study used Shuttle Radar Topography Mission (SRTM) data with a 30-metre resolution. The mean-shift segmentation was applied repeatedly with different parameters to capture the variability of the alluvial fans. Subsequently, an SVM was used to classify the already grouped objects. The results showed that this approach, called Object-Based Morphometric Analysis (OBMA), achieved good results. It was measured using fuzzy membership values derived from the SVM classification to select the most appropriate segmentation for each identified alluvial fan.

Babic et al. (2021) also modelled and classified alluvial fans using DEMs and ML techniques. The study focused on torrential alluvial fans in Slovenia, identifying seven main geomorphometric parameters: mean hinterland slope, mean torrent slope, Melton basin roughness number, relief ratio, the ratio between fan area and hinterland area, Melton alluvial fan number, and mean fan slope. By comparing five ML methods, including Random Forest (RF), Genetic Programming, SVM, Neural Network, and a hybrid Euler graph method, the researchers demonstrated these approaches' effectiveness in automatically classifying alluvial fans prone to debris flows. The study utilised data from various satellite image sources, such as ASTER, GeoEye, Ikonos, WorldView, ALOS, and SPOT Image, with resolutions ranging from 2 metres (WorldView) to 30 metres (ASTER and SRTM). The results, validated with empirical data, showed that Genetic Programming performed the best in classification.

Rabanaque et al. (2021) conducted a large-scale hydro morphological analysis of ephemeral streams using ML algorithms, specifically SVM and RF, to segment and classify river channels and associated fluvial forms. The input variables included active channel width, valley floor width, slope gradient, route distance, and specific stream power, along with remote sensing data from Sentinel-2 spectral bands (RGB, NIR1, SWIR1, SWIR2) and spectral indices as the Normalised Difference Vegetation Index (NDVI), Green Red Vegetation Index (GRVI), and Normalised Difference Water Index (NDWI), as well as textural indices like variance, correlation, contrast, entropy, second moment, mean, and dissimilarity. LiDAR data from the Plan Nacional de Ortofotografía

Aérea (PNOA-LiDAR) Project were used to create a DEM with a resolution of 25 m, which was subsequently resampled to 10 m using bilinear interpolation. This technique calculates the value of a new pixel as a weighted average of the four nearest original pixels. The accuracy of the models was assessed using the confusion matrix, predictive accuracy, and Cohen's Kappa index, with the SVM achieving an average accuracy of 0.87 and Kappa of 0.84. In contrast, the RF achieved an average accuracy of 0.85 and Kappa of 0.81.

The history of alluvium identification reflects a continuous evolution from observational methods to increasingly sophisticated techniques. Advancements in computational power have driven this progression, the need to better understand geomorphological processes, and the imperative to manage natural resources more effectively.

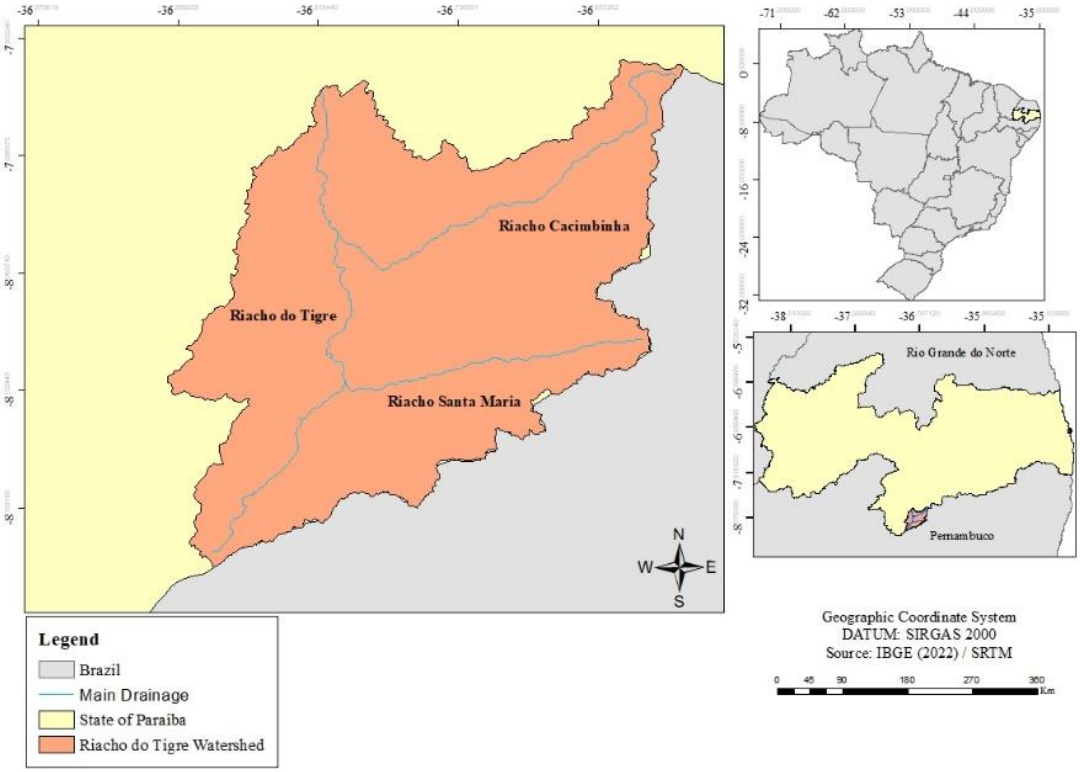
MATERIALS AND METHODS

The study area is the Riacho do Tigre watershed, located in the city of São João do Tigre, in the State of Paraíba. Approximately 560 km² is situated on the Borborema Plateau, characterised by the predominance of crystalline rocks, influencing the region's relief. The altitude of the watershed ranges approximately from 500 to 1100 metres. The slope varies between 0 and 114%. The climate is tropical semi-arid, with an average annual rainfall of 431.8 mm (Silva, 2017). These watershed characteristics are directly related to the landscape formation in the area. Higher regions are sediment formation zones, while lower regions are deposition areas for these elements transported by rainwater. This process is crucial for the formation of the alluvial areas.

The Riacho do Tigre watershed comprises three main streams (Figure 2). The principal stream is Riacho do Tigre, with altitudes ranging from 500 to 800 meters and an approximate length of 40 km. The Cacimbinha stream has a length of 32 km and an altitude variation between 550 and 750 meters. Additionally, the watershed includes the Santa Maria stream, which is 21 km long with altitudes varying between 600 and 950 meters.

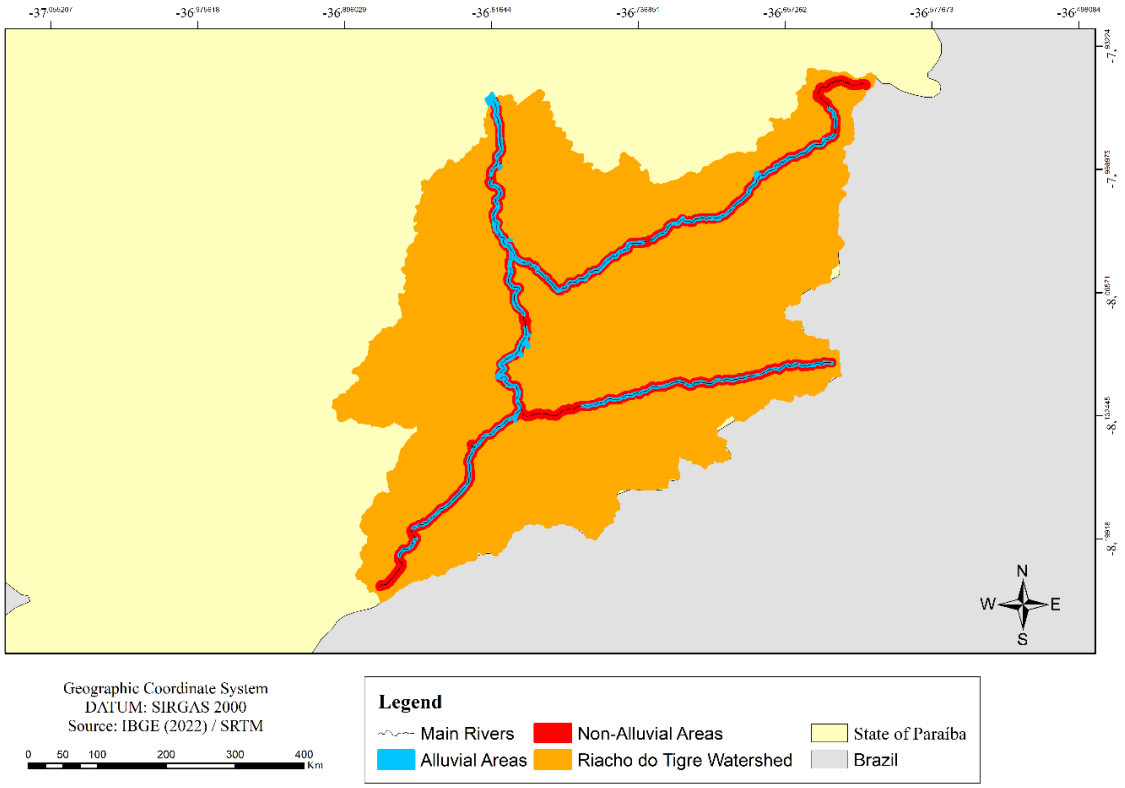
Figure 3 presents the sections identified as alluvial areas within the Riacho do Tigre watershed. These regions, highlighted in blue, denote areas where sediment deposition occurs as a result of fluvial dynamics. The map also delineates non-alluvial areas, the primary river courses, and the watershed boundary.

Figure 2 - Riacho do Tigre Watershed



Drafting: Authors (2024).

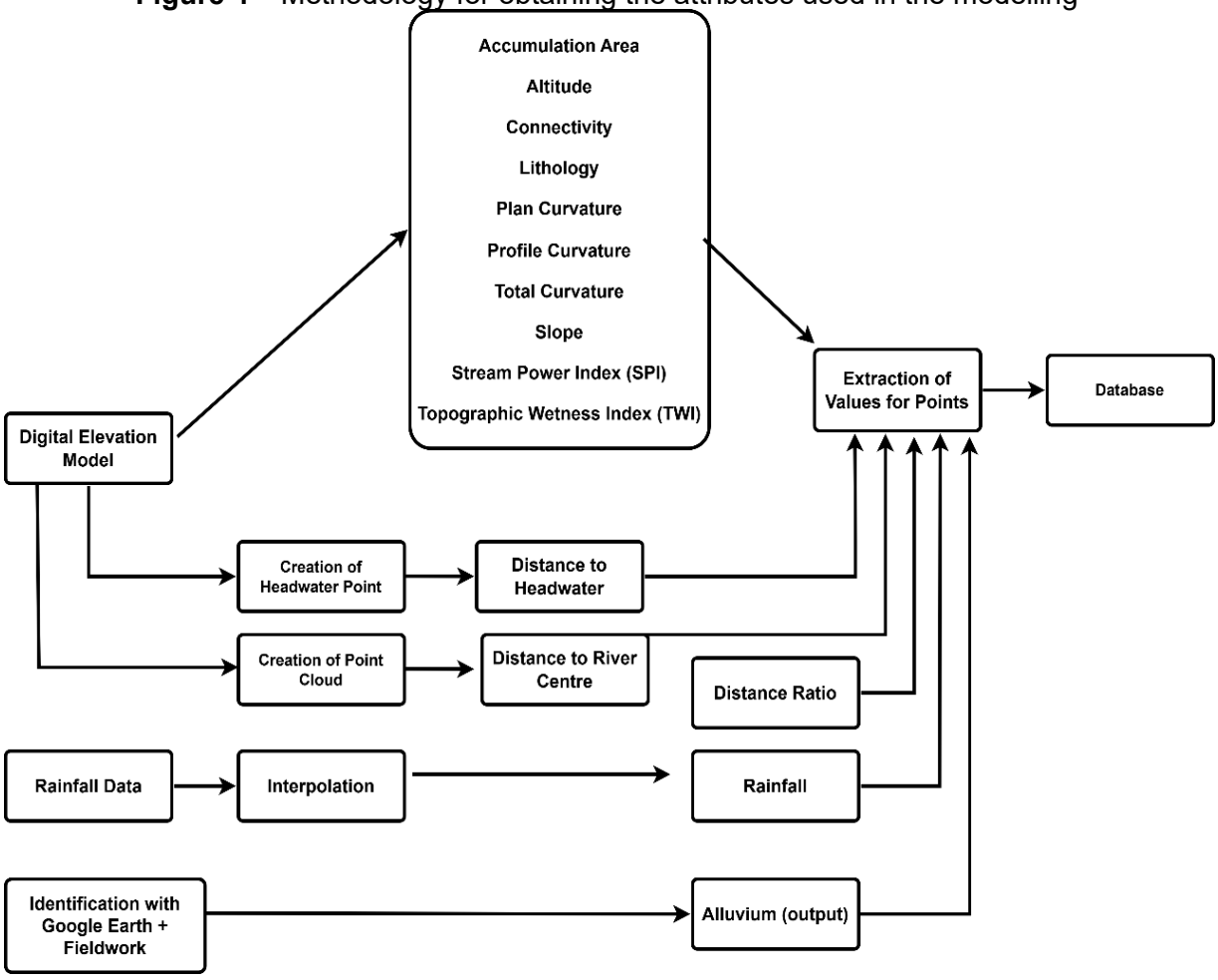
Figure 3 – Spatial Distribution of Alluvial Areas in the Riacho do Tigre Watershed



Drafting: Authors (2024)

The acquisition of the fifteen attributes that comprise the database was conducted using ArcMap software, version 10.8, a Geographic Information System (GIS) tool. Through this tool, essential spatial and geographical analyses were carried out for the collection of variables. The first step of the work involved processing SRTM images with a resolution of 30 meters. The flowchart in Figure 4 outlines the steps to obtain the database.

Figure 4 – Methodology for obtaining the attributes used in the modelling



Drafting: Authors (2024)

Along the three main rivers of the Riacho do Tigre basin, a 210-meter buffer zone was created on each side from the centre of the channel. Within this polygon, 1,458,886 points were distributed. Creating the point cloud was a crucial methodological step for obtaining the fourteen input and output variables. The table below characterises each of the attributes used in the modelling.

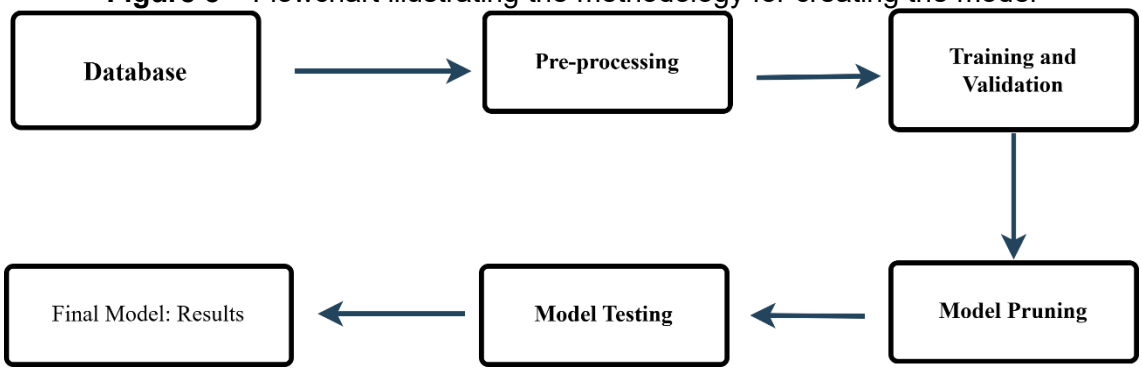
Box 1 – Modelling Variables

Attribute	Description
Altitude	The elevation relative to sea level is measured in metres for each point.
Accumulation Area	Indicates the accumulation area from the headwater to the point, measured in square kilometres.
Index of Connectivity	The index measures, on a pixel scale, the connectivity of a given point with other parts of the basin. The index ranges from [-∞, +∞].
Plan Curvature	Represents the curvature of the terrain in the horizontal plane.
Profile Curvature	Represents the curvature of the terrain in the vertical plane.
Total Curvature	Represents the curvature of the terrain by combining the two previous curvatures.
Slope	Measures the inclination, in degrees, of the terrain relative to the horizontal.
Distance to River Centre	Measures the distance of each point, in metres, to the centre of the channel.
Distance to Headwaters	Measures the distance of each point, in metres, to the headwater of the river basin.
Lithology	A categorical variable that extracted the lithological type under each point: 1. Granitoids, 2. Metagranitoids, 3. Metamorphic Complexes, and 4. Alluvial Areas. Data were obtained from maps of the Geological Survey of Brazil (SGB/CPRM).
Rainfall	Based on data from five pluviometric stations (São João do Tigre, Camalau, São Sebastião do Umbuzeiro, Jataúba, and Poções), interpolation was conducted to measure the amount of rainfall (mm) for each point. The data used in the interpolation were obtained from the Agência Executiva de Gestão das Águas do Estado da Paraíba (AESA) and corresponded to the average precipitation over the last 30 years.
Distance Ratio	A ratio with the distance to the river centre as the numerator. The denominator is the sum of the distance from the river centre to the channel margins.
SPI	Stream Power Index measures the erosive power of running water on the terrain.
TWI	Topographic Wetness Index, a topographic index that indicates soil moisture.
Alluvium (Output)	Alluvium is the output variable. Points identified as alluvial areas were recorded as 1, while non-alluvial areas were recorded as 0.

Drafting: Authors (2024)

The development of this stage of the work was carried out using MATLAB software version R2024a. The Statistics and Machine Learning Toolbox package (Version 24.1, R2024a) was also utilised. The flowchart in Figure 5 illustrates the methodological steps adopted for data modelling and the DT creation.

Figure 5 – Flowchart illustrating the methodology for creating the model



Drafting: Authors (2024)

Preprocessing was conducted from the data importation, which involved removing some attributes that were not part of the modelling and partitioning the set of variables into input and output data for training and testing.

Following this step, the model was trained using the fitctree function in MATLAB. This function employs the Gini Index as a criterion to split the decision tree nodes. Model validation was performed through k-fold cross-validation using the crossval function. This method allowed for the assessment of model performance by dividing the dataset into k equal parts, training the model on k -1 parts, and testing it on the remaining part. The value of k was iterated from five to one hundred, with the average accuracy recorded for each iteration. Additionally, the bootstrap method was applied to calculate confidence intervals for the average accuracies, using 2000 samples for each value of k, providing a significant estimate of the variability of these indices.

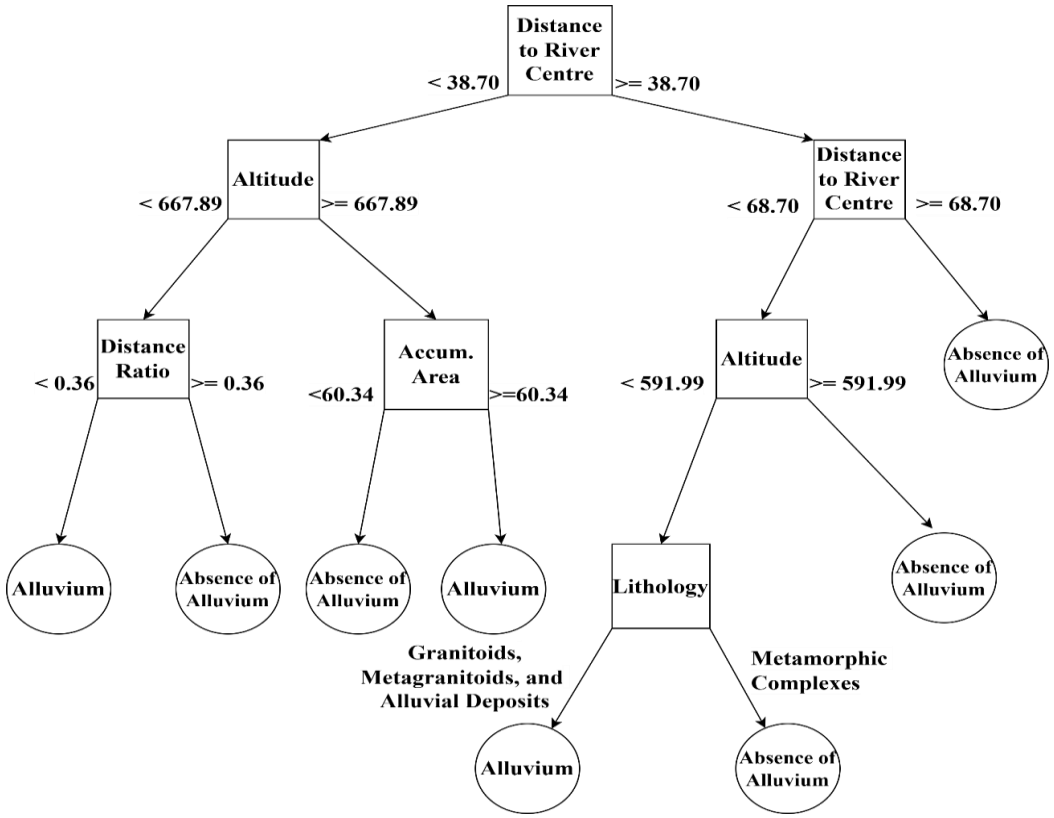
After validating and testing the model, pruning was performed using the minimum cost-complexity pruning technique. Various alpha values were tested to identify the pruning that resulted in the best accuracy with the fewest number of nodes in the tree. The best configuration was determined based on the obtained accuracy, resulting in a simpler and more effective tree. The results and evaluation metrics for the model construction, including pruning details, will be presented in the next chapter.

RESULTS

Figure 6 depicts the classification tree that represents the final and most accurate model obtained to detect the presence of alluvial areas. The decision to use a DT in this research was due to its comprehensibility, providing a clear and accessible interpretation

of the results. The DT illustrates the hierarchical structure used to classify alluvial areas based on various input variables. The tree's root node uses the variable "Distance to River Centre", with a cut-off point of 38.7 metres. If the distance to the channel is less than this value, the tree branches to the left; otherwise, it branches to the right. On the left branch, if the "Altitude" is less than 667.89 metres, the decision then depends on the "Distance Ratio," where values less than 0.36 indicate class 1 (alluvial area).

Figure 6 – Decision Tree Used for the Classification of Alluvial Areas



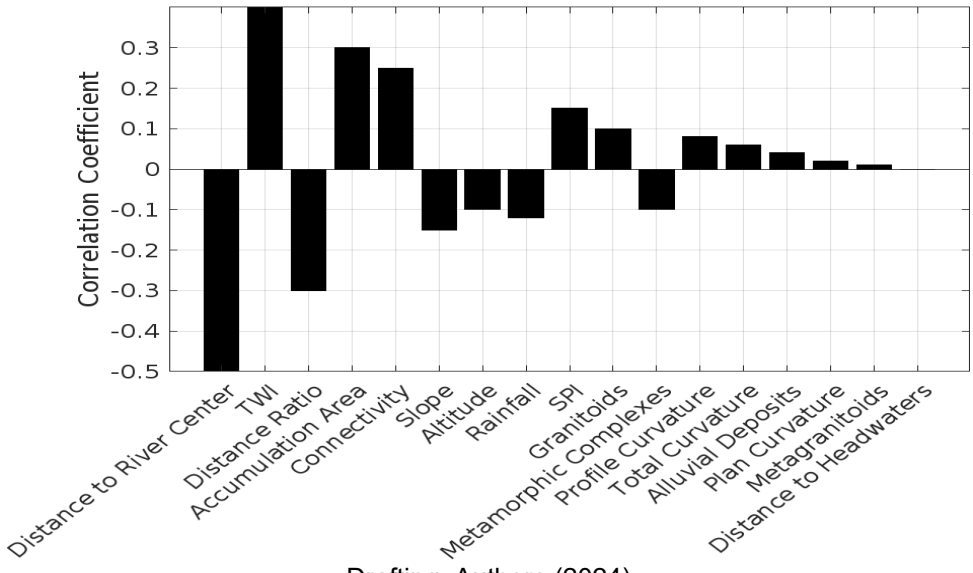
Drafting: Authors (2024)

In contrast, values equal to or greater indicate class 0 (non-alluvial areas). If the "Altitude" is greater than or equal to 667.89 metres, the tree evaluates the "Accumulation Area"; areas with accumulation less than 60.34 square kilometres are classified as class 0, and those equal to or greater as class 1. On the right branch of the tree, if the "Distance to the River Centre" is greater than or equal to 38.70 and less than 68.70 metres, the decision is based on "Altitude" at 591.99 metres. The classification for altitudes less than this value depends on "Lithology"; if the lithology is categorised as Granitoids, Metagranitoids, or an Alluvial Deposit area, the class is 1.

Otherwise, the class is 0. Altitudes equal to or greater than 591.99 metres are classified as class 0. Finally, distances to the channel equal to or greater than 68.7 metres are always classified as class 0.

Correlation analysis is a statistical technique used to measure the strength and direction of the linear relationship between two variables based on Pearson's correlation coefficient (R). The variables that showed the most significant correlations were (Figure 7): Distance to the River Centre, with a correlation of -0.56, exhibited the strongest negative relationship, indicating that shorter distances to the river centre are associated with a higher likelihood of alluvial deposits. TWI (Topographic Wetness Index) had a positive correlation of 0.41, suggesting that areas with higher topographic moisture are more likely to exhibit alluvial deposits. Distance Ratio, with a correlation of -0.34, indicates that smaller distance ratio values are associated with a higher probability of alluvial deposits. The variable Accumulation Area, with a correlation of 0.32, suggests that areas with more significant water accumulation are more likely to have alluvial deposits. Connectivity showed a correlation of 0.29, indicating that hydrologically more connected areas are associated with a higher likelihood of alluvial deposits. Slope, with a negative correlation of -0.29, indicates that steeper terrains are less likely to have alluvial deposits. Other variables such as Altitude, Rainfall, SPI (Stream Power Index), and various geological formations (Granitic, Metamorphic Complexes, Alluvial Deposits and Metagranitoids) showed lower intensity correlations with the output Variable.

Figure 7 – Correlation between Input and Output Variables



Drafting: Authors (2024)

The performance results of the classification model, as presented in Table 2, indicate that the accuracy, precision, recall, and other evaluation metrics achieved consistent values, highlighting the model's effectiveness in detecting alluvial areas.

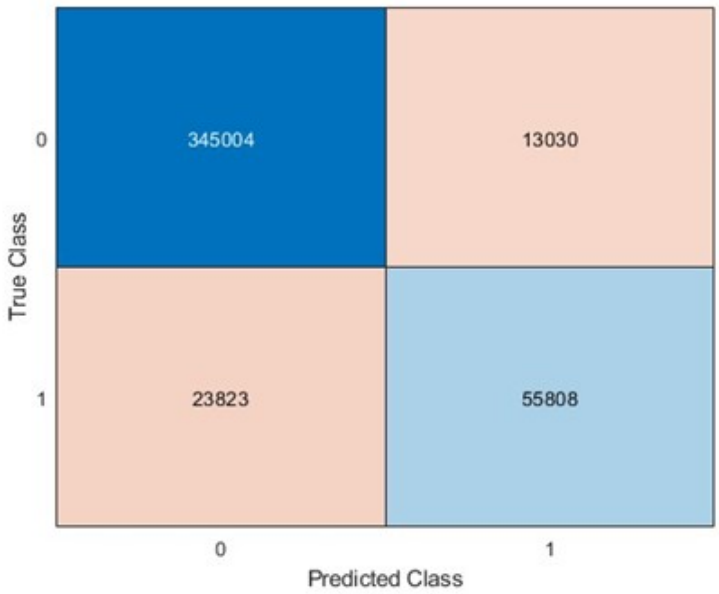
Box 2 – Model Performance Results

Metric	Value
Accuracy (The ratio of correct predictions to the total number of predictions)	0.92
Precision (The ratio of true positive predictions to the total predicted positive cases)	0.81
Recall (The ratio of true positive predictions to the total actual positive cases)	0.70
F1-Score (The harmonic mean of precision and recall)	0.75
Error Rate (The ratio of incorrect predictions to the total number of predictions)	0.08
Area Under ROC (The area under the ROC curve representing model discrimination)	0.91

Drafting: Authors (2024)

The accuracy measures the total correct predictions concerning the total number of points. The model achieved an accuracy of 0.92 during the testing phase. The confusion matrix of the model (Figure 8) indicates that the Precision, which is the number of correct predictions for alluvial areas in relation to the total points classified as alluvial, was 0.81.

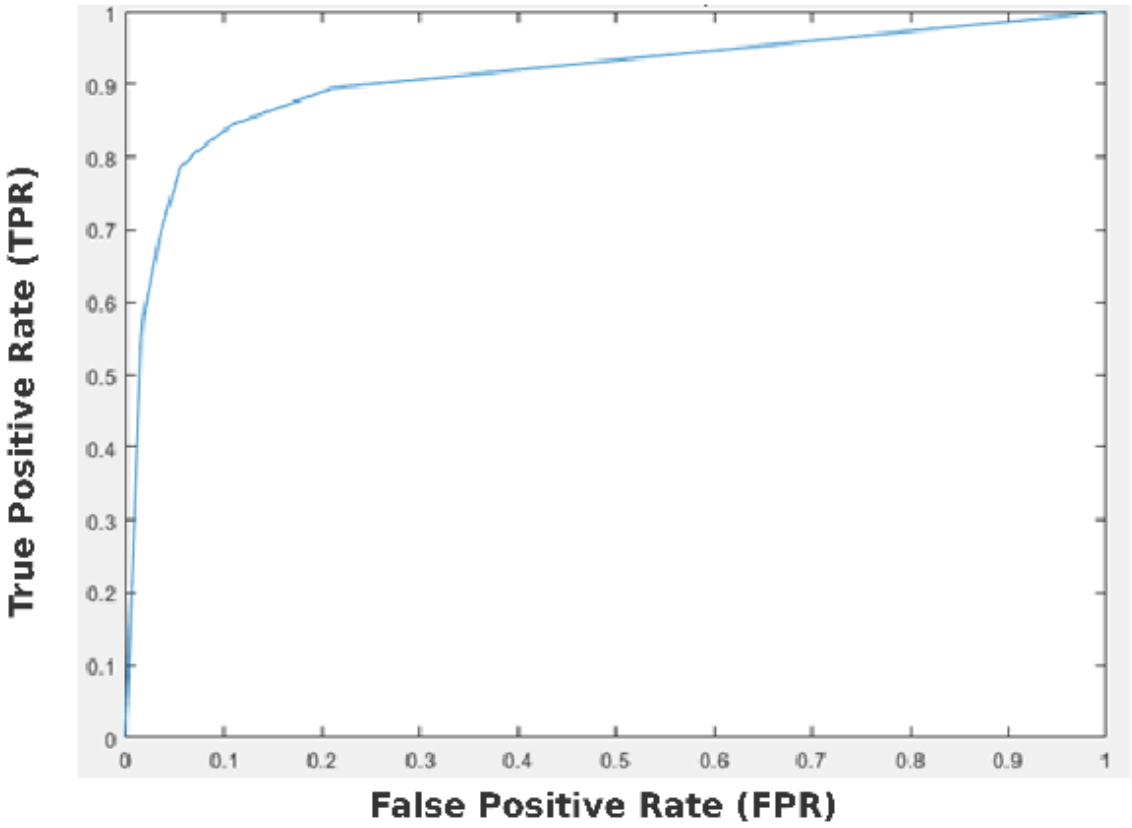
Figure 8 – Confusion Matrix for Model Performance Evaluation



Drafting: Authors (2024)

Recall, which measures the number of correctly identified alluvial areas with the total alluvial points, was 0.70. The F1-Score, the harmonic mean between precision and recall, was 0.75. The error rate found by the model, which is the ratio of the sum of incorrect evaluations to the total evaluations, was 0.084. The ROC curve (Receiver Operating Characteristic) is shown below. The area under the ROC curve was 0.91, indicating good discrimination between positive and negative classes.

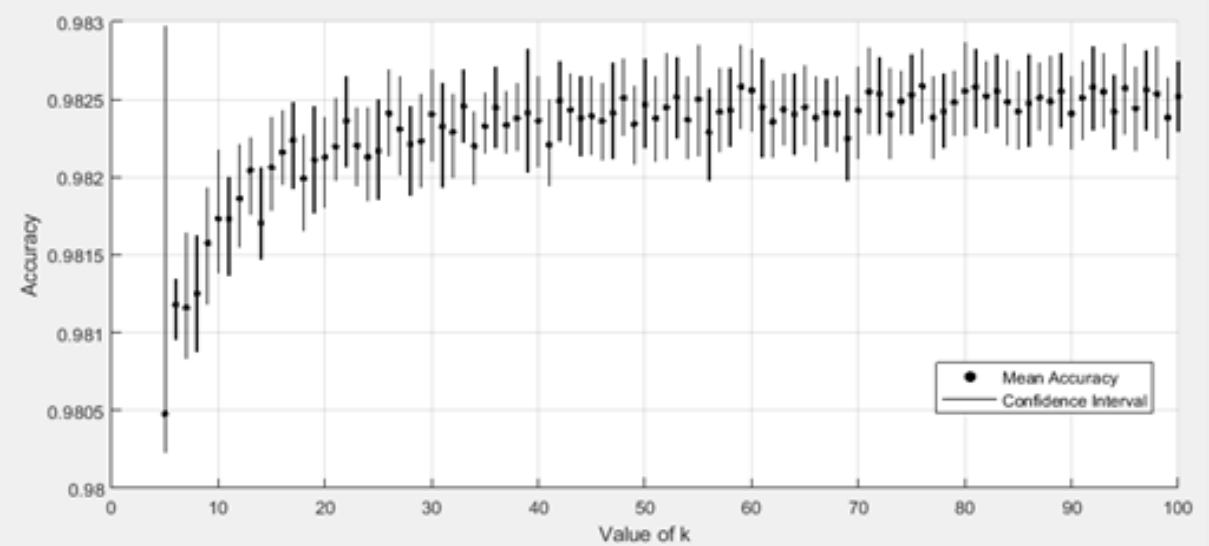
Figure 9 – ROC Curve for Model Performance Evaluation



Drafting: Authors (2024)

The DT model was validated using k-fold cross-validation and Bootstrap resampling. Initially, cross-validation was applied with different values of k (ranging from 5 to 100) to calculate the mean accuracy. It was observed that the mean accuracy remained consistently high, indicating strong model performance.

Figure 10 – Mean Accuracy and Confidence Intervals for k-Fold Cross-Validation



Drafting: Authors (2024)

To ensure the precision of our estimates, we used bootstrapping with 2000 samples to calculate the confidence intervals of accuracy for each value of k. The results in Figure 10 presented narrow confidence intervals, suggesting that the estimates are robust and precise. After validation, we selected the model corresponding to the k value with the highest average accuracy. This model, generated with k = 76, was subsequently used in the pruning process to refine the DT further.

In this study, we pruned the DT using the cost complexity technique to enhance the model's generalisation and avoid overfitting. We defined a range of values for the alpha parameter, which controls the tree's complexity. Using an iterative approach, we varied alpha across 500 points. For each alpha value, the tree was pruned, and both the accuracy and the number of remaining nodes in the pruned tree were calculated. This approach allowed us to assess the impact of different pruning levels on the model's performance. After analysing the results, we selected the alpha value (0.0021) that offered an optimal balance between accuracy and model simplicity.

DISCUSSION

The analysis of the importance of predictors, measured by the reduction of the Gini impurity criterion attributed, reveals the most relevant factors for the DT model. Distance to Channel" emerged as the most significant attribute, substantially contributing to the model's accuracy. It indicates that areas closer to channels have a significantly

higher probability of presenting alluvium due to water flow dynamics and sediment deposition. "Altitude" was identified as the second most important predictor, suggesting that the elevation of the terrain strongly influences the occurrence of alluvium, possibly by affecting the speed and volume of surface runoff. "Distance Ratio," although of lesser importance, is still relevant, indicating that the relationship between different geomorphological distances can impact the deposition of alluvium. "Accumulation Area" also proved significant, suggesting that areas accumulating more water have a greater propensity to develop alluvium due to the increased capacity for sediment transport. Finally, "Lithology," representing the types of rocks and soils present, stood out as a relevant factor, possibly influencing soil permeability and resistance to erosion. The low correlation between the output variable and the categorical variable "Alluvial Deposits" suggests a potential need to revise and update the data from the Geological Survey of Brazil (SGRiB/CPRM). This discrepancy indicates that the current representation of alluvial deposits by the SGB may not accurately reflect the lithological reality of the studied areas.

When comparing the results of this study with those of Pipaud and Lehmkuhl (2017), Babic et al. (2021), and Rabanaque et al. (2021), clear similarities emerge in the use of geomorphological and hydrological variables across all studies. Pipaud and Lehmkuhl (2017) utilised parameters such as curvature and slope to identify alluvial fans. Babic et al. (2021) focused on geomorphometric variables, such as the relationship between area and slope to classify torrential fans. In contrast, Rabanaque et al. (2021) integrated distance-related variables, including channel width and route distance, into their hydromorphological analyses of ephemeral streams.

In this study, the most significant variables for the decision tree model that identified alluvial deposits were 'Distance to Channel', 'Altitude', 'Distance Ratio', 'Accumulation Area', and 'Lithology'. While there are similarities in using parameters such as curvature and slope, this study differs by emphasising distance and accumulation area variables, which are particularly relevant for characterising alluvial deposits in ephemeral streams. This shift in focus highlights the difference between the types of formations studied: whereas previous research concentrated on alluvial fans, this work focuses on identifying alluvial deposits along channels. As a result, variables such as 'Distance to Channel' and 'Accumulation Area' are identified as key factors in the model.

These observations highlight the importance of specific geomorphological and hydrological parameters in identifying different types of alluvial formations. While Pipaud and Lehmkuhl (2017) and Babic et al. (2021) demonstrated the relevance of curvature and slope in the classification of alluvial fans, this study emphasises the crucial role of distance-related variables, such as 'Distance to Channel' and 'Accumulation Area', in the detection of alluvial deposits. This contrast in parameters reflects the different scales and geomorphological processes involved. Alluvial fans are typically larger-scale formations influenced by broader topographical features, which explains the focus on slope and curvature in previous studies. On the other hand, alluvial deposits, particularly in semi-arid environments, are more sensitive to hydrological dynamics and proximity to water channels. It justifies the prioritisation of distance and accumulation variables in this study.

The analysis of the indices indicates that the model is accurate. However, there is still ample room for improving its recall to identify more areas that are truly alluvial correctly. Considering the inherent limitations of using images with a spatial resolution of 30 metres, such as those from the SRTM, the results must be discussed. This resolution presents significant limitations in detecting alluvial areas as it may not capture fine details of alluvial features, such as small sediment deposits and subtle variations in terrain morphology. Additionally, the 30-metre resolution may not be sufficient to accurately pinpoint the exact location of alluvial areas due to the varying sizes of these features. Various input data, such as altitude, slope, and land cover, are restricted to this resolution, resulting in a simplified and generalised landscape view. This simplification can lead to inaccurate classifications and hinder the precise identification of the boundaries of alluvial areas.

The 30-metre images are the only ones available for the watershed studied in the present research. Although they are helpful for large-scale analyses and provide an essential overview, studies focused on detecting alluvial areas may require higher spatial resolution for more detailed and accurate results. Various studies have used higher-resolution data to improve the accuracy of identifying geomorphological features. For example, Pipaud and Lehmkuhl (2017) employed DEMs with a 10-metre spatial resolution, which allowed for a more detailed representation of terrain features. Babic et al. (2021) utilised a similar resolution, contributing to a more refined analysis of torrential fans. On the other hand, Rabanaque et al. (2021) used a 5-metre DEM, which resulted in even greater precision in their hydromorphological analysis of ephemeral streams.

These variations in spatial resolution had a noticeable impact on the results: studies that used finer resolutions could capture smaller-scale features and produce more accurate classifications of alluvial formations.

CONCLUSION

There is a global increase in concern for water resources and a growing interest in studies aimed at identifying groundwater. This research, conducted in a semi-arid area of northeastern Brazil, utilised a DT to classify points in the three main channels of the basin as alluvial areas.

The model achieved an accuracy of 0.92. The area under the ROC curve was 0.91, indicating good discrimination between positive and negative classes. The analysis of the indices suggests that the model is accurate. However, there is still much room for improvement in its recall, allowing it to correctly identify more truly alluvial areas.

Furthermore, the results obtained in this study demonstrate that the application of ML techniques, specifically DT, is effective in detecting alluvial deposits. The analysis of predictor importance revealed that variables such as "Distance to the River Centre" and "Altitude" are crucial for the model's accuracy. Cross-validation and Bootstrap confirmed the robustness and precision of the model's estimates. However, the 30-metre spatial resolution of the images used presents significant limitations in capturing details of alluvial features. Therefore, higher-resolution data will be used in future studies to improve the identification and mapping of such areas.

REFERÊNCIAS

ALI, R. et al. Effectiveness of machine learning ensemble models in assessing groundwater potential in Lidder watershed, India. Acta Geophysica, v. 72, p. 2843-2856, 2023.

AL-JUAIDI, F. et al. Merged remotely sensed data for geomorphological investigations in deserts: examples from central Saudi Arabia. The Geographical Journal, v. 169, p. 117-130, 2003.

AMIT, R. et al. Soils as a tool for estimating ages of Quaternary fault scarps in a hyperarid environment: the southern Arava valley, the Dead Sea Rift, Israel. Catena, v. 28, p. 21-45, 1996.

ARDABILI, S. et al. Deep learning and machine learning in hydrological processes climate change and earth systems: a systematic review. Engineering for Sustainable Future, v. 101, p. 55-62, 2020.

BABIC, M. et al. Modeling and Classification of Alluvial Fans with DEMs and Machine Learning Methods: A Case Study of Slovenian Torrential Fans. Remote Sensing, v. 13, p. 1-18, 2021.

BRAGA, R. A. P. et al. Águas de Areia. 1. ed. Recife: Clã, 2016.

BÚRQUEZ, A. et al. Human-made small reservoirs alter dryland hydrological connectivity. Science of the Total Environment, v. 947 (10): 174673, 2024.

CAVALLI, M. et al. The effectiveness of airborne LiDAR data in the recognition of channel-bed morphology. Catena, v. 73, p. 249-260, 2008.

CAVALLI, M.; MARCHI, L. Characterisation of the surface morphology of an alpine alluvial fan using airborne LiDAR. Natural Hazards and Earth System Sciences, v. 8, p. 323-333, 2008.

CERVI, F.; TAZIOLI, A. Quantifying Streambed Dispersion in an Alluvial Fan Facing the Northern Italian Apennines: Implications for Groundwater Management of Vulnerable Aquifers. Hydrology, v. 8 (3):118, 2021.

CROUVI, O. et al. Quantitative mapping of arid alluvial fan surfaces using field spectrometer and hyperspectral remote sensing. Remote Sensing of Environment, v. 104, p. 103-117, 2006.

DÍAZ-ALCAIDE, S.; MARTÍNEZ-SANTOS, P. Review: Advances in groundwater potential mapping. Hydrogeology Journal, v. 27, p. 2307-2324, 2019.

EL BILALI, A et al. Comparing four machine learning model performances in forecasting the alluvial aquifer level in a semi-arid region. Journal of African Earth Sciences, v. 181:104244, 2021.

EFRON, B.; TIBSHIRANI, R. J. An introduction to the bootstrap. Chapman & Hall/CRC, 1993.

FARR, T. G.; CHADWICK, O. Geomorphic processes and remote sensing signatures of alluvial fans in the Kun Lun Mountains, China. Journal of Geophysical Research, v. 101, p. 23091-23100, 1996.

FRANKEL, K. L.; DOLAN, J. F. Characterizing arid region alluvial fan surface roughness with airborne laser swath mapping digital topographic data. Journal of Geophysical Research, v. 112, p. 1-14, 2007.

GABER, A. et al. Textural and Compositional Characterisation of Wadi Feiran Deposits, Sinai Peninsula, Egypt, Using Radarsat-1, PALSAR, SRTM and ETM+ Data. Remote Sensing, v. 2, p. 52-75, 2010.

GAFFOOR, Z. et al. A Comparison of Ensemble and Deep Learning Algorithms to Model Groundwater Levels in a Data-Scarce Aquifer of Southern Africa. Hydrology, v. 9, n. 7, 2022.

GILLESPIE, A. R. et al. Mapping Alluvial Fans in Death Valley, California using Multichannel Thermal Infrared Images. Geophysical Research Letters, v. 11, p. 1153-1156, 1984.

GOORABI, A. et al. Semi-automated method for the mapping of alluvial fans from DEM. Earth Science Informatics, v. 14, p. 1447-1466, 2021.

GHOLAMI, V. et al. Prediction of annual groundwater depletion: An investigation of natural and anthropogenic influences. Journal of Earth System Science, v. 132, 160, 2023.

HASTIE, T et al. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer, 2009.

HERTZ, G. et al. Estimating the age of desert alluvial surfaces with spaceborne radar data. Remote Sensing of Environment, v. 184, p. 288 – 301. 2016.

HOHENTHAL, J. et al. Laser scanning applications in fluvial studies. Progress in Physical Geography, v. 35, p. 782-809, 2011.

IACOBUCCI, G. Land-surface quantitative analysis for mapping and deciphering the construction processes of piedmont alluvial fans in the Anti-Lebanon Mountains. Geomorphology, v. 453, 109148, 2024.

JASECHKO, S. et al. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature, v. 625, n. 7996, p. 715–721, 2024.

KAYHOMAYOON, Z. et al. How does a combination of numerical modeling, clustering, artificial intelligence, and evolutionary algorithms perform to predict regional groundwater levels? Computers and Electronics in Agriculture, v. 203, 107482, 2022.

LIMA, I. Inteligência Artificial. 1. ed. São Paulo: Grupo GEN, 2014.

LUIZ, T. B. P. Previsão de Dados de Água Subterrânea utilizando modelos baseados em Aprendizado de Máquina. Tese (Doutorado em Engenharia Civil) – Santa Maria: UFSM, 2022.

MARSLAND, S. Machine Learning: an algorithmic Perspective. Boca Raton: Taylor & Francis Group, LLC, 2015.

MARTÍNEZ-SANTOS, P.; RENARD, P. Mapping Groundwater Potential Through an Ensemble of Big Data Methods. Groundwater, v. 58, n. 4, p. 583–597, 2020.

MCLEOD, J. S. et al. Landscapes on the edge: River intermittency in a warming world. Geology, v. 52, n. 7, p. 512-516, 2024.

MITCHELL, T. M. Machine Learning. McGraw-Hill Science/Engineering/Math, 1997.

MUÑOZ-CARPENA, R. et al. Convergence of mechanistic modeling and artificial intelligence. Hydrologic science and engineering. Plos Water, v. 2, 59, 2023.

NGUYEN, P. T. et al. Improvement of credal decision trees using ensemble frameworks for groundwater potential modeling. Sustainability, v. 12, n. 7, 2622, 2020.

PIPAUD, I.; LEHMKUHL, F. Object-based delineation and classification of alluvial fans by application of mean-shift segmentation and support vector machines. Geomorphology, v. 293, p. 178-200, 2017.

QUINLAN, J. R. Induction of Decision Trees. Machine Learning, v. 1, n. 1, p. 81-106, 1986.

RABANAQUE, M. P. et al. Basin-wide hydromorphological analysis of ephemeral streams using machine learning algorithms. Earth Surface Processes and Landforms, v. 47, p. 1-17, 2021.

REGMI, N. R. et al. Mapping Quaternary alluvial fans in the southwestern United States based on multiparameter surface roughness of lidar topographic data. Journal of Geophysical Research, v. 119, p. 12-27, 2014.

RITCHIE, H.; EISMA, J. A.; PARKER, A. Sand dams as a potential solution to rural water security in drylands: Existing research and future opportunities. Frontiers in Water, v. 3, 31, 2021.

RUSSELL, S. J.; NORVIG, P. Inteligência Artificial: Uma Abordagem Moderna. Rio de Janeiro: Grupo GEN, 2022.

SEIFU, T. K. et al. Application of advanced machine learning algorithms and geospatial techniques for groundwater potential zone mapping in Gambela Plain, Ethiopia. Hydrology Research, v. 54, n. 10, p. 1246–1266, 2023.

SHAKYA, et al. Groundwater level prediction with machine learning for the Vidisha district, a semi-arid region of Central India. Groundwater for Sustainable Development, v. 19, 100825, 2022.

SILVA, A. F. P.; SOUZA, J. P. Análise da qualidade da água nos aquíferos aluviais da bacia Riacho do Tigre-PB: Uma abordagem hidrológica em ambientes fluviais semiáridos no Brasil. Revista de Geografia Norte Grande, n. 84, p. 323–336, 2023.

SILVA, A. F. P. L.; SOUZA, J. O. P. Caracterização Hidrossedimentológica dos trechos fluviais da bacia do Riacho do Tigre – PB. Caminhos de Geografia, v. 18, n. 63, p. 57–89, 2017.

SOUZA, J. O. P.; ALMEIDA, J. D. M. Processos fluviais em terras secas: uma revisão. Revista OKARA: Geografia em Debate, v. 9, p. 108–122, 2015.

SRIVASTAVA, D. K.; SHUKLA, A.; JEMNI, D. Prediction of Ground Water Level in Rajasthan State Using Machine Learning. Procedia Computer Science, v. 218, p. 1702-1711, 2023.

STALEY, D. M. et al. Surficial patterns of debris flow deposition on alluvial fans in Death Valley, CA using airborne laser swath mapping data. Geomorphology, v. 74, p. 152-163, 2005.

TAO, H. et al. Groundwater level prediction using machine learning models: A comprehensive review. Neurocomputing, v. 489, p. 271-308, 2022.

TAYER, T. C. et al. Identifying intermittent river sections with similar hydrology using remotely sensed metrics. Journal of Hydrology, v. 626, 130266, 2023.

UC-CASTILLO, J. L. et al. A systematic review and meta-analysis of groundwater level forecasting with machine learning techniques: Status and future directions. Environmental Modelling and Software, v. 168, 105788, 2023.

VADIATI, M. et al. Application of artificial intelligence models for prediction of groundwater level fluctuations: case study (Tehran-Karaj alluvial aquifer). Environmental Monitoring and Assessment, v. 194, 619, 2022.

ZARESEFAT, M.; DERAKHSHANI, R. Revolutionizing Groundwater Management with Hybrid AI Models: A Practical Review. Water, v. 15, 1750, 2023.

.

.