Data-Driven Insights into Deforestation: Predictive Modeling in Colombian Regions

Alvaro Hernán Alarcón-López[1,*], Ixent Galpin[1]

[1]Universidad de Bogota-Jorge Tadeo Lozano, Bogota, Colombia

Abstract

Deforestation is a critical problem that affects biodiversity, climate patterns, and water quality. This article presents a predictive model for the deforestation rate in the regions of Colombia using the CRISP-DM methodology. Historical data from 2015 to 2022 from the Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM) were used. Correlation analyses were performed and models were trained using Random Forest to obtain the predictor variables: deforested and regenerated area, stable forest area, net difference in forest cover area, and change in forest cover area. Additionally, the departments of Atlántico, Sucre, Santander, and Meta were identified as the regions with the highest deforestation rates. In the prediction process, linear regression models showed the highest accuracy, with an R^2 of 1.00. Finally, the importance of segmenting and analyzing data by region to obtain accurate predictions and take effective corrective measures is highlighted.

Keywords

Deforestation, CRISP-DM methodology, Random Forest, Annual Deforestation Rate, Machine learning

1. Introduction

Deforestation is a global issue that has garnered significant attention from scientists and environmentalists due to its numerous adverse effects on the environment [1, 2]. The loss of biodiversity and alterations in climate patterns are among the most devastating consequences, profoundly impacting ecosystem health and human well-being. Additionally, deforestation has disrupted watershed dynamics and aquatic ecosystems, contributing to the deterioration of water quality and ecological habitats. While deforestation in Colombia is reportedly at an all-time low [3], it is particularly serious for the country due to its rich biodiversity and vital ecosystems, which are crucial for maintaining global climate stability, water cycles, and the livelihoods of indigenous and local communities. To tackle these challenges, various technologies and analytical methods have been developed to identify and predict deforested areas, providing essential data for devising effective conservation and reforestation strategies [4, 5].

In recent years, analytical and classification models have become crucial for understanding and predicting areas affected by deforestation. For instance, drone imagery has been instrumental in accurately identifying deforested areas and monitoring forest regeneration over time. These advanced technologies have significantly enhanced the precision of deforestation detection and have optimized reforestation efforts [6].

Similarly, the application of machine learning for classifying satellite images has enabled the precise identification of areas affected by forest fires and other disturbances. Moreover, studies utilizing classification techniques to analyze deforestation trends underscore the importance of continuous monitoring and conservation efforts. These tools and methods have also been employed to predict future trends, aiding in the development of effective strategies for ecosystem preservation.

This paper presents a predictive model for deforestation rates in various regions of Colombia using the CRISP-DM methodology a process model for data mining successful in research and development [7]. The objective is to provide a technological tool for the early implementation of corrective actions.

ICAIW 2024: Workshops at the 7th International Conference on Applied Informatics 2024, October 24–26, 2024, Viña del Mar, Chile
*Corresponding author.
alvaroh.alarconl@utadeo.edu.co (A. H. Alarcón-López); ixent@utadeo.edu.co (I. Galpin)
0000-0003-4703-1907 (A. H. Alarcón-López); 0000-0001-7020-6328 (I. Galpin)

The structure of the work is as follows: Section 2 discusses the impacts of deforestation, including biodiversity loss, decreased water availability, and increased soil erosion. Section 3 reviews previous research on the prediction and mitigation of deforestation. In Section 4, the departments in Colombia with the highest deforestation rates based on forest area are identified, and the data distribution is analyzed. Section 5 establishes the correlation between predictor and target variables and quantifies their importance. Section 6 identifies predictor variables by department and generates prediction models. Section 7 evaluates the performance of each model using appropriate metrics for regression problems. Finally, Section 9 presents the study's findings and provides a comprehensive overview of the research.

2. Deforestation

Deforestation of forests can lead to a series of long-term, observable problems and consequences, ultimately resulting in serious environmental issues. One of the primary impacts is the reduction of biodiversity, alteration of ecosystem functioning, and modification of carbon dynamics [8, 9]. Furthermore, the loss of natural habitats for numerous species can cause the extinction of endemic flora and fauna, thereby diminishing the unique biodiversity of each region.

Another impact directly related to the reduction of forested areas is the significant increase in global temperatures. This is due to the rise in carbon dioxide levels released into the atmosphere from deforestation, which contributes to global warming and disrupts climate patterns [10]. As a result, these changes intensify extreme weather events, adversely affecting human communities and ecosystems that depend on a stable climate for their survival and well-being.

Furthermore, deforestation impacts the availability of water in watersheds and alters its flow and distribution. These changes have significant implications for terrestrial hydrological systems and the ecosystems that depend on them [11]. Additionally, the loss of forest cover can lead to greater variability in water flow, resulting in more frequent and severe droughts and floods.

Another direct consequence of deforestation is the increase in soil erosion rates and the disruption of nutrient and water cycles, which adversely affect the livelihoods of local communities. This degradation reduces soil quality and its capacity to support agriculture. This situation underscores the importance of implementing sustainable forestry, agricultural, and livestock practices to mitigate negative impacts and safeguard the natural resources essential for human sustenance.

On the other hand, in South America, deforestation has additional consequences related to the reduction of glacier recharge, which feeds rivers and returns water to the Amazon. This situation poses a serious threat to the future of agriculture in various natural regions, as it leads to atypical occurrences of droughts and floods, thereby increasing the likelihood of environmental and forest disasters [12].

In the regions of Colombia, deforestation leads to several critical issues, including reduced biodiversity, altered ecosystem functioning, significant contributions to global temperature rise due to increased atmospheric carbon dioxide, and disrupted climate patterns [13]. It also affects water availability in watersheds, altering flow and distribution, and increases soil erosion rates, disrupting nutrient and water cycles, thus adversely impacting the survival of living organisms. Furthermore, agricultural activities, such as livestock farming, have exacerbated deforestation rates through unsustainable practices like burning forest areas to create grazing land for cattle [14].

3. Related Work

Analysis and classification models are paramount for understanding and predicting deforestation and its associated problems. Numerous studies have been conducted on this topic. One example is the use of convolutional neural networks (CNN) [15] for multiclass semantic segmentation, which enables the identification of deforested areas from drone images. The goal of this research was to selectively distribute seeds and monitor forest regeneration over time [6].

Another approach is the application of machine learning to satellite image classification to identify areas affected by wildfires. This method has demonstrated high potential for accurately classifying

such images, utilizing metrics such as precision and average success rate [4]. Additionally, the use of pre-trained convolutional neural networks (CNNs), combined with clustering algorithms like K-Means, has enabled the precise identification of damaged forest areas. This provides an effective solution for labeling satellite data, supporting rapid reforestation efforts [16].

In another study by Kani *et al.*, RF classification was used to analyze deforestation trends, revealing a gradual decrease in forest areas over the years. This study underscores the importance of continuous monitoring and conservation efforts, emphasizing the need for immediate actions to prevent further loss of forest areas [5]. By leveraging these models and analytical techniques, it is possible not only to accurately identify deforested areas but also to predict future trends. This capability contributes to the development of effective reforestation and ecosystem preservation strategies, enabling more sustainable and effective forest management.

In the review conducted, no investigative studies were found that attempted to develop prediction models based on time series analysis for Colombia. Therefore, this work is novel in this field as it undertakes a distinctive approach compared to the existing research. This study aims to develop a methodology to predict the deforestation rate in the regions of Colombia, to enable early corrective actions. To achieve this, the well-established CRISP-DM methodology, specialized in data mining, is employed. This methodology is based on a hierarchical model distributed across different development stages: business understanding, data understanding, data preparation, modeling, and evaluation [7].

4. Data Understanding

The dataset used in this study comprises 561 records across 33 of 52 departments in Colombia, with each department represented by 17 records. These records detail deforestation rates and related factors, including variables such as forest area (SFA), deforested area (DA), and deforestation rate (ADR), among others, over various time periods. The dataset includes key departments like Amazonas, Atlántico, Sucre, Santander, and Meta, enabling a comprehensive analysis of deforestation trends. The even distribution of data across these regions is essential for evaluating the robustness of the models and the complexity of the analysis, ensuring that localized deforestation patterns are accurately captured and modeled.

In this phase, the available and necessary resources are evaluated, and the objective of data mining is determined. Data from secondary sources are collected and described, and their quality is verified by statistical analysis, determining attributes and correlations [7]. The data consists of a historical record of environmental statistics provided by the Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM). Specifically, two datasets were used: 'Change in the area covered by natural forest according to Department Consolidated results between 1990-2022' and 'Annual deforestation rate according to Department Consolidated results between 1990-2022'[1]. Complementary variables from these datasets are used, with the understanding that they have identical time series and that both contain segmented data from the 33 departments. The analysis of historical annual data from 2005 to 2022 for each of the 33 departments of Colombia is carried out to observe trends and changes over time. Table 1 presents the data dictionary used for the analysis.

To develop the prediction model, the annual deforestation rate (ADR) is selected as the dependent variable. Due to the data dispersion, the models are developed by regions. Consequently, the independent (predictor) variables are defined by the department.

Due to the segmentation of data by departments, it is essential to determine which departments exhibit a higher deforestation rate relative to the proportion of stable forest area in each region. To achieve this, a new dataset is generated that presents the calculated averages for each of the variables by department. Additionally, a column is added to establish the relationship between NDAC (net deforestation, calculated as the difference between DA and RA) and the stable forest area variable (SFA).

From this initial analysis, it was determined that the departments of Atlántico, Sucre, Santander, and Meta exhibited the highest NDAC/SFA ratios. Consequently, these regions had higher deforestation

[1] http://www.ideam.gov.co/web/ecosistemas/bosques-y-recurso-forestal

Table 1
Data dictionary

Variable	Type	Description
SFA	Decimal numeric	Stable forest area (ha)
DA	Decimal numeric	Deforested area (ha)
RA	Decimal numeric	Regenerated area (ha)
AWI	Decimal numeric	Area without information (ha)
PAWI	Decimal numeric	Proportion of area without information (%)
ADR	Decimal numeric	Annual deforestation rate (%)
NDAC	Decimal numeric	Net difference of the area covered by forest period t1 : t2
CFA	Decimal numeric	Change in forest area

rates relative to the amount of forest in their territory. For this study, these four departments were selected, and a summary of the analysis is presented in Table 2.

Table 2
Higher NDAC/SFA ratios

Department	NDAC/SFA
Atlántico	-0.027893
Sucre	-0.023736
Santander	-0.009413
Meta	-0.008593

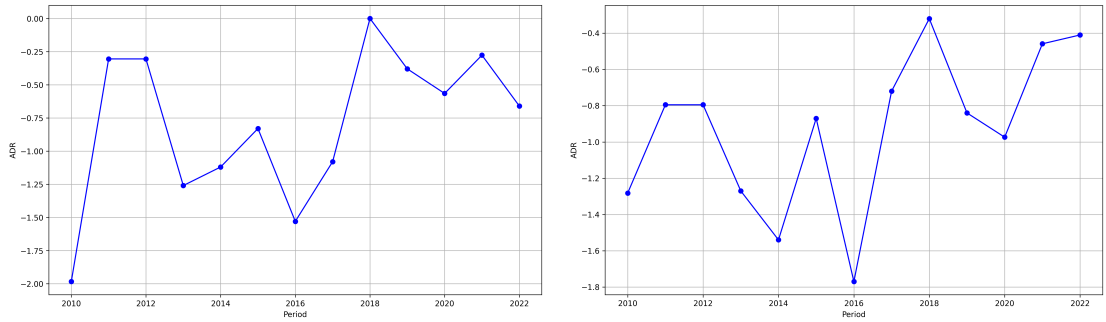
Figure 1 presents an ADR graph for the departments of Atlántico, Sucre, Santander, and Meta, highlighting distinct deforestation patterns. In Atlántico, the trend is fluctuating with significant peaks and abrupt drops, indicating periods of intensive deforestation followed by recovery. Sucre exhibits notable variability, with significant increases and sudden decreases in ADR. Santander displays high variability characterized by multiple peaks and valleys. In Meta, a decreasing trend in ADR is observed, suggesting the possible effectiveness of conservation policies; however, sporadic peaks still occur.

To visualize the distribution of the annual deforestation rate for these four departments, the original dataset was filtered, and a boxplot was generated. This boxplot allows for a visual comparison of the variability and distribution of the annual deforestation rate among the selected departments. Figure 2 reveals that none of the distributions contain outliers for the analyzed variable. Additionally, it is noted that the department of Atlántico has the highest median deforestation rate, while Santander has the lowest median.

5. Data Preparation

Data selection is performed by defining specific inclusion and exclusion criteria for the IDEAM dataset, using various methods described in the corresponding section [7]. Initially, records with missing values (NA) are eliminated and some variables are converted from decimals to integers to ensure consistency in the analysis. Once the data are clean, we proceed to the construction of derived attributes, such as the NDAC/SFA ratio, which could serve as an additional predictor variable in the model.

The Random Forest model is selected at this initial stage because of its ability to handle large numbers of variables and its ability to identify the most relevant features among them. Unlike other models, Random Forest is not affected by multicollinearity and can efficiently handle data with high dimensionality. This approach allows exploring the dataset in depth, identifying precisely which variables have the greatest impact on the prediction of the annual rate of deforestation.



(a) Atlántico (b) Sucre

(c) Santander (d) Meta

Figure 1: ADR for departments with highest NDAC/ SFA ratio

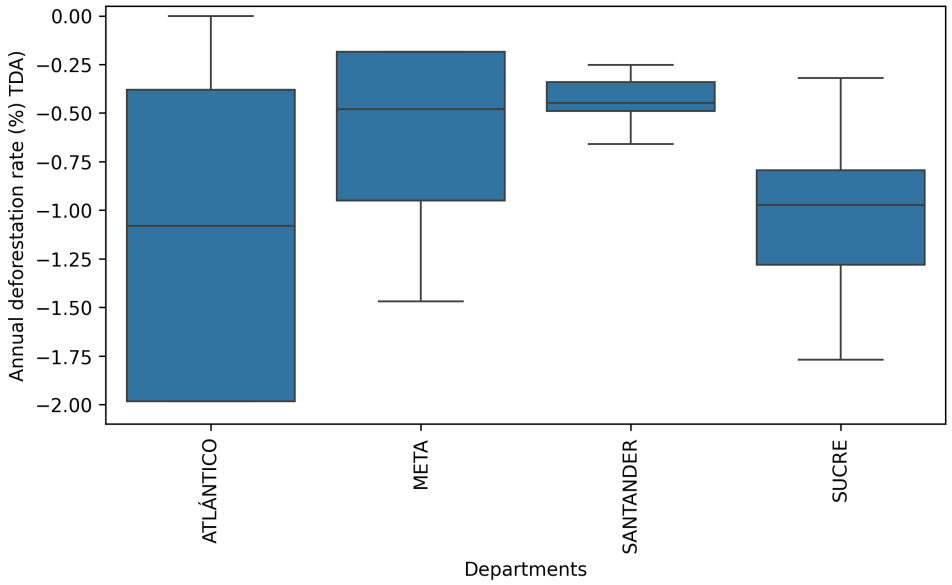
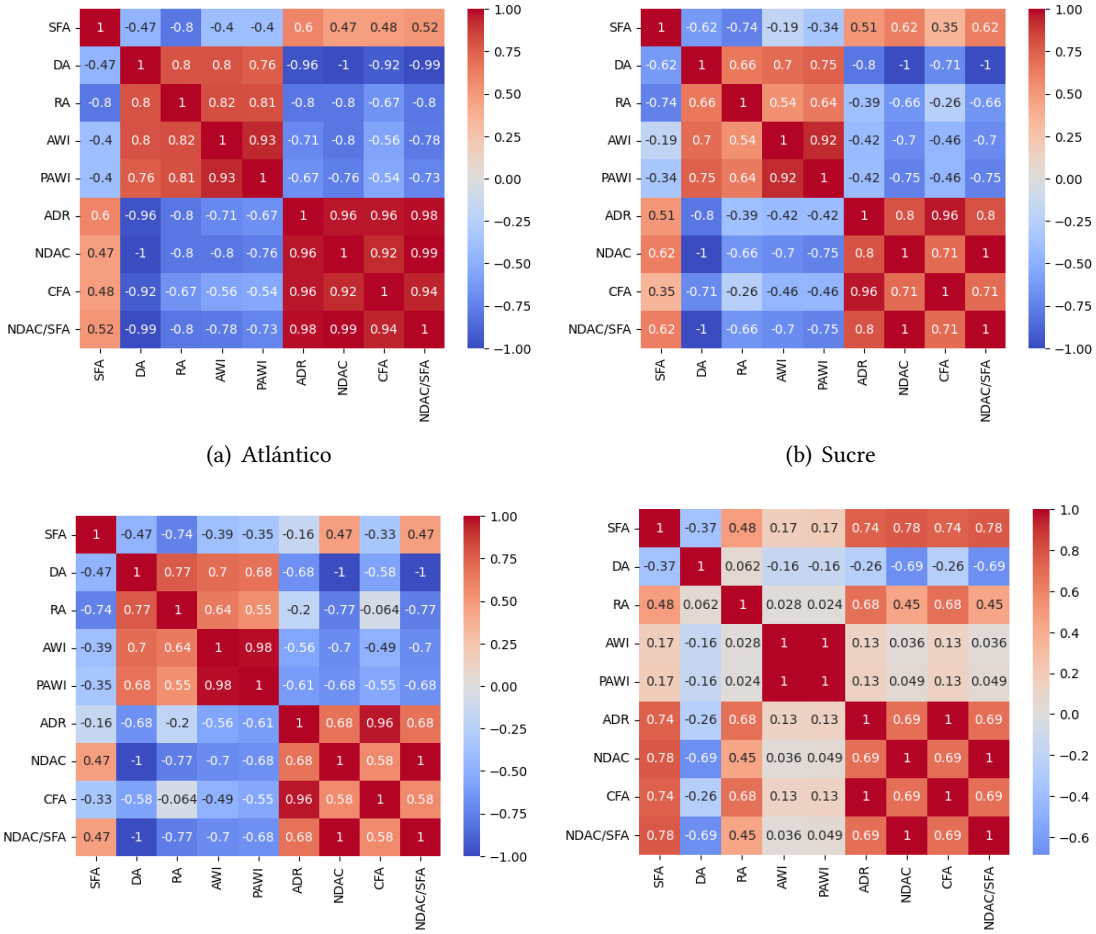


Figure 2: Deforestation annual rate distribution

Furthermore, the Random Forest model provides a valuable measure of the importance of features, which facilitates the identification of the most significant variables for prediction. This capability is critical in high-dimensional studies, such as deforestation analysis, where it is crucial to determine which variables have the greatest impact on the results. By prioritizing the most relevant features, Random Forest helps reduce the risk of overfitting and improves the predictive capability of the model [17]. This measure of importance not only guides model building but also provides a deeper understanding of the factors driving changes in forest area.

In this context, correlation plots were generated for all variables, as well as box plots and scatter plots of the annual rate of deforestation for each of the four departments with the highest NDAC/SFA ratio. These graphs allowed a clear and detailed visualization of the relationship between the selected variables and annual deforestation, which confirmed the relevance of the chosen characteristics. Using these visual methods in combination with the feature importance measure provided by Random Forest ensures that the model is based on the most robust and reliable predictors available, thus optimizing its ability to make accurate and useful predictions for deforestation management. Figure 3 shows the correlation matrix for each of the selected departments.



(a) Atlántico (b) Sucre

(c) Santander (d) Meta

Figure 3: Spearman correlation matrix to departments with highest NDAC/ SFA ratio

5.1. Significant Correlations

In the department of Atlántico, the variables DA, RA, NDAC, NDAC/SFA, and CFA exhibit a high correlation with ADR (annual deforestation rate (%)). The p-value for each of these variables in relation to ADR was determined using the Mann-Whitney statistical test and was found to be less than 0.05 for all of them. This result rejects the null hypothesis, indicating that these variables could be strong predictors. The calculated values are presented in Table 3.

In the Sucre region, the variables DA, NDAC, CFA, and NDAC/CFA show a high correlation with ADR(annual deforestation rate). The p-value calculations using the Mann-Whitney test yielded values lower than 0.05 for each of these variables, leading to the rejection of the null hypothesis. Therefore, it is concluded that these variables could be strong predictors. The calculated values are presented in Table 4.

Table 3
P-value for variables with higher correlation - Atlántico

Variables	Mann-Whitney (p-value)
DA	1,71E+07
RA	2,57E+07
NDAC	1,04E+09
CFA	1,04E+09
NDAC/SFA	1,04E+09

Table 4
P-value for variables with higher correlation - Sucre

Variables	Mann-Whitney (p-value)
DA	8,57E+05
NDAC	8,57E+05
CFA	8,57E+05
NDAC/SFA	8,57E+05

Additionally, in the department of Santander, the variables DA, NDAC, CFA, and NDAC/CFA exhibit a high correlation with ADR (annual deforestation rate). The p-value calculations using the Mann-Whitney test yield values below 0.05 for each of these variables, leading to the rejection of the null hypothesis. This suggests that these variables could be strong predictors. The calculated values are presented in Table 5.

Table 5
P-value for variables with higher correlation - Santander

Variables	Mann-Whitney (p-value)
DA	8,57E+05
NDAC	8,57E+05
CFA	8,57E+05
NDAC/SFA	8,57E+05

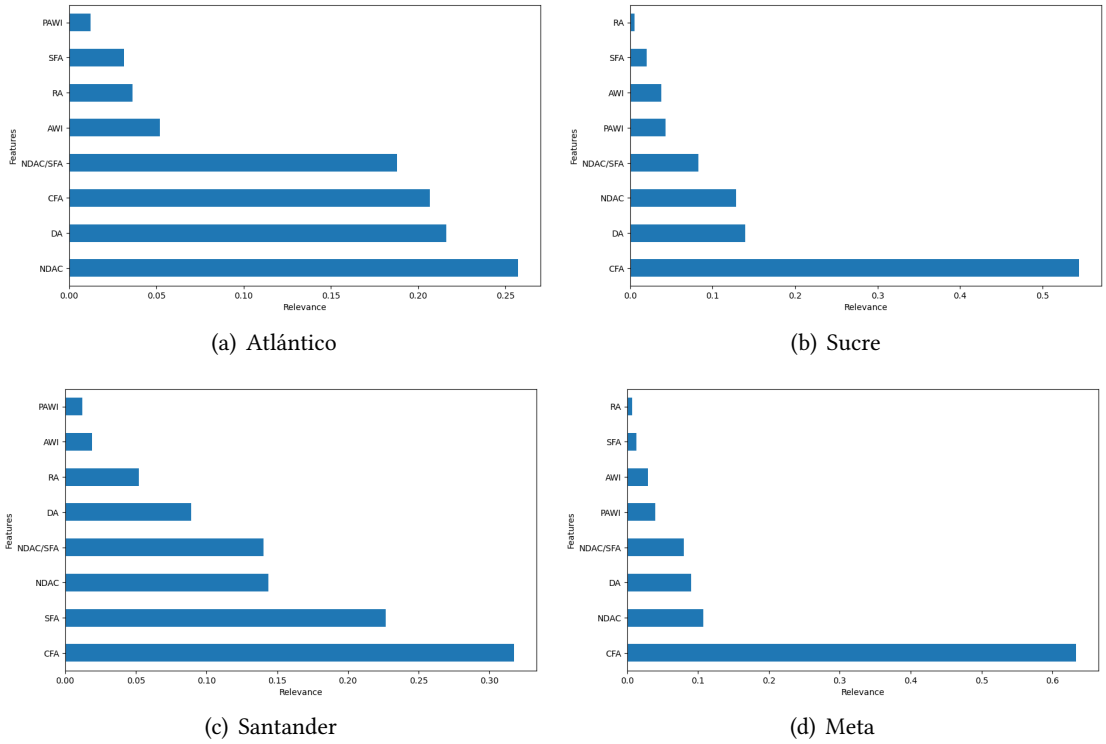
In the Meta region, the variables SFA, RA, NDAC, and NDAC/CFA show a high correlation with ADR (annual deforestation rate). The p-value calculations using the Mann-Whitney test yield values below 0.05 for each of these variables, leading to the rejection of the null hypothesis and indicating that these variables could be strong predictors. The calculated values are presented in Table 6.

Table 6
P-value for variables with higher correlation - Meta

Variables	Mann-Whitney (p-value)
SFA	8,57E+05
RA	8,57E+05
NDAC	8,57E+05
NDAC/SFA	8,57E+05

5.2. Relevance features

To confirm and quantify the importance of the variables, a Random Forest model is trained, as this algorithm allows the results of multiple decision trees to be combined to reduce the risk of overfitting and improve the generalization of the model [18]. This is in contrast to linear and logistic regression, which do not adequately capture the complex and non-linear relationships between variables, and other models such as SVM and neural networks, which can be more costly and difficult to interpret. Therefore, a ratio of 80% training data, 20% test data and the evaluation metric used was RSME. Predictor variables were defined as: SFA, DA, RA, AWI, PAWI, NDAC, CFA, and NDAC/SCBE and the target variable: ADR. In this way, the aim was to determine the importance of these characteristics for each of the selected departments. selected departments. The results obtained can be seen in Figure 4.



(a) Atlántico (b) Sucre

(c) Santander (d) Meta

Figure 4: Importance ratio of the features

For the Atlántico department, the Random Forest model achieved an RMSE of 0.0692, indicating good accuracy due to the relatively small average error. Regarding feature importance, NDAC is identified as the most important variable, contributing 25.72%. DA also shows significant importance with 21.6%, followed by CFA at 20.68% and the NDAC/SFA ratio at 18.78%. The remaining features each contribute less than 5%.

For the department of Sucre, the Random Forest model achieved an RMSE of 0.0904. The most important variable is CFA, contributing 54.4%. SD follows with 13.94%, NDAC accounts for 12.83%, and the NDAC/SFA ratio contributes 8.23%, while the remaining features each have an importance of less than 5%. For the department of Santander, the RMSE is 0.0227. Here, CFA is again the most important variable, contributing 63.32%. NDAC follows with 10.74%, DA accounts for 9.01%, and the NDAC/SFA ratio contributes 7.98%, with the other features each contributing less than 5%.

Finally, for the Meta department, the Random Forest model achieved an RMSE of 0.0772. The most important variable is CFA, contributing 31.73%. SFA follows with 22.64%, NDAC accounts for 14.35%, the NDAC/SFA ratio contributes 14.03%, and DA represents 8.9%. The remaining features each have an importance of less than 5%.

6. Modeling

The technique selection and model development phase is essential in the predictive modeling process, as it determines the tools and methods that will be used to analyze the data. Therefore, to approach the problem of forest area change and deforestation rate using the data provided by IDEAM, the technique that best suits the nature of the problem and the quality of the available data must be chosen. It is essential to consider that the selected methods must be able to handle both linear and nonlinear relationships present in the data [19], which will allow capturing the complex patterns inherent to the deforestation process.

In the case of IDEAM data, models such as linear regression, decision trees, SVM (Support Vector Machines), and random forest are selected to predict the ADR (Annual Deforestation Rate). These models are chosen because of their ability to combine linear and nonlinear techniques, which allows for capturing diverse and complex patterns in the data.

Each selected technique offers particular advantages that make it suitable for this type of analysis. Linear regression is used to identify simple relationships between variables, providing a clear basis for understanding how certain factors influence the rate of deforestation. On the other hand, decision trees and random forest models are effective for capturing more complex interactions between variables, being especially useful when working with data that exhibit nonlinear relationships. In addition, SVM is especially valuable in high-dimensional scenarios, where the number of variables can complicate other simpler methods.

Although there are more advanced techniques, such as neural networks, they are not used in this case due to their high computational demands and the need for large volumes of data for effective training. Neural networks are powerful and can capture very complex patterns, but their implementation requires significant resources and a larger data set than was available. Therefore, it was decided to use models that offer a balance between predictive capability and computational efficiency. The source code used to develop these models in Python is available on GitHub[2], allowing other researchers to reproduce the results or adapt the techniques to their datasets.

6.1. Atlántico Region Model

The predictive model for the annual rate of deforestation in the Atlantico region is developed after a detailed analysis using the correlation index, the p-value, and the significance key features index. These analyses allow the identification of the most relevant variables for the model. In this scenario, it is identified that the variables NDAC, NDAC/SCBE, CFA, DA, and RA are the most effective predictors for modeling the annual rate of deforestation in the data from the Department of Atlántico. These variables reflect a strong correlation with deforestation, suggesting that the combination of anomalous climatic factors and the current forest situation is critical for predicting changes in forest cover in the region.

6.2. Sucre Region Model

In the Sucre region, data analysis showed that several variables are essential for predicting the annual rate of deforestation. The correlation index, p-value, and importance ranking of the characteristics determined that the variables NDAC, NDAC/SFA, CFA, and DA should be used as predictors. The integration of these variables into the analysis provides a robust framework that facilitates not only the accurate prediction of the annual rate of deforestation but also strategic and informed forest management decisions. This model reflects the specific realities of the region, providing a useful tool for adaptation processes to environmental and social changes, and strengthening the strategies for the conservation and sustainable use of forests in Sucre.

[2] https://github.com/AlvaroHernan/DeforestationPredictive

6.3. Santander Region Model

In the case of the department of Santander, the variables that most influence the annual rate of deforestation were identified, thanks to the analysis of the correlation indexes, the p-value, and the classification of the importance of the characteristics. The study concluded that the variables NDAC, NDAC/SFA, CFA, and DA are the most relevant for the predictive model. Their inclusion in the model provides a solid basis for understanding the underlying drivers of deforestation, which is essential for developing effective conservation strategies.

6.4. Meta Region Model

The model designed for the department of Meta is based on a comprehensive analysis that has identified the variables NDAC, SFA, NDAC/SFA, CFA, and DA as essential for predicting the annual rate of deforestation. This model not only provides an accurate prediction of changes in forest area but also acts as a valuable resource for informed decision-making in natural resource management, offering a crucial tool for the formulation of adaptive and effective conservation strategies in the Meta region.

7. Results

In the evaluation phase, model results are compared using R^2, MSE, and MAE metrics to assess model accuracy. To enhance the robustness of the evaluation, cross-validation is used. This technique divides the data into multiple subsets, or folds, ensuring that the model is trained and tested on different partitions of the data. A common approach is 5-fold cross-validation, where the data is split into five parts, training the model on four parts and testing it on the remaining one, repeating the process five times. This provides a more comprehensive assessment of the model's performance, in particular with regards to its generalization capabilities. Metrics such as R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE) are computed for each fold, and their average values are used to determine the overall precision of the models, offering a more reliable evaluation than a simple train test split. This section presents the interpretation of the obtained results to extract relevant and significant conclusions. The results for each model by the department are presented in Tables 7–10.

Table 7
Model results - Atlántico

Model	R^2	MSE	MAE
Linear Regression	1.00	0.00	0.00
Decision Tree	0.76	0.03	0.11
SVM	0.27	0.11	0.27
Random Forest Regressor	0.91	0.02	0.10

Table 8
Model results - Sucre

Model	R^2	MSE	MAE
Linear Regression	1.00	0.00	0.00
Decision Tree	0.80	0.01	0.06
SVM	-0.16	0.09	0.23
Random Forest Regressor	0.71	0.02	0.08

The results indicated that the linear regression model was the most accurate in predicting deforestation rates across all the analyzed departments: Atlántico, Sucre, Santander, and Meta. The model achieved

Table 9
Model results - Santander

Model	R^2	MSE	MAE
Linear Regression	1.00	0.00	0.00
Decision Tree	0.89	0.00	0.02
SVM	0.47	0.01	0.07
Random Forest Regressor	0.96	0.00	0.01

Table 10
Model results - Meta

Model	R^2	MSE	MAE
Linear Regression	1.00	0.00	0.00
Decision Tree	0.91	0.01	0.09
SVM	0.93	0.01	0.08
Random Forest Regressor	0.89	0.01	0.07

R^2 values close to 1.0, reflecting its high accuracy in forecasting deforestation patterns in these regions and demonstrating a superior ability to explain the variability in deforestation rates.

In addition to linear regression, the random forest model also showed a competitive performance, especially in the departments of Atlántico and Meta, with R^2 above 0.91. This model is known for its ability to capture complex interactions between variables, which makes it particularly useful in contexts where deforestation patterns are influenced by multiple interrelated factors. Although the random forest did not outperform the linear regression model in terms of R^2, its results were close enough to consider it a robust alternative, especially in scenarios where it is desirable to minimize the risk of overfitting.

On the other hand, decision tree models and support vector machines (SVM) presented lower performance compared to linear regression and random forest. In the case of the decision tree model, the R^2 values ranged from 0.76 to 0.91, indicating that, although effective, its ability to predict accurately is lower than that of the aforementioned models. The SVM model, although useful in specific contexts, showed the greatest limitations, with R^2 ranging from -0.16 to 0.93, suggesting that it may not be the best choice for this type of analysis in regions with complex and highly variable data such as deforestation.

The analysis of the mean squared error (MSE) and mean absolute error (MAE) supported the conclusions obtained from R^2. In all departments, the linear regression not only presented the lowest MSE and MAE values but also maintained remarkable consistency among the different data sets. This fact reinforces the idea that linear regression is not only accurate but also stable in its performance, which is crucial for the implementation of policies based on its predictions. The performance of the models was carefully interpreted to draw relevant conclusions.

Importantly, the superiority of linear regression could be due to the linear nature of the underlying relationships between predictor variables and deforestation rate. However, the slight variability in the performance of the models in different departments also underscores the importance of considering specific regional characteristics when selecting the most appropriate model.

In summary, the linear regression model emerged as the most effective tool for predicting deforestation rates in the departments evaluated, providing highly accurate and reliable predictions. The random forest stood out as a robust alternative, especially in more complex scenarios. The results obtained underline the importance of a regionalized approach to predictive modeling,

8. Future Work

The incorporation of variables such as temperature, humidity, and forest type into predictive models is crucial for improving the accuracy of predictions, but it faces significant challenges in terms of obtaining and managing these data. The quality and availability of accurate and updated information on these variables can be difficult to guarantee, especially in remote regions or areas with limited infrastructure for environmental data collection. The limited number of meteorological stations in certain forests and variability in collection methods can generate inconsistencies that affect the accuracy of the model. In addition, available historical data may not cover long enough periods to capture long-term trends, limiting the model's predictive capability.

Another significant limitation lies in the temporal and spatial resolution of the data. In many instances, climatic and forest information is available at a broad scale, making it challenging to conduct the detailed local analyses required for accurate deforestation modeling. This lack of data granularity can lead to models that fail to capture critical variations within regions, thus reducing the effectiveness of conservation strategies based on these predictions. Additionally, the model's performance was affected by the inherent variability of the data and inconsistencies across departments. To manage these inconsistencies, a feature selection process using Random Forest was employed, enabling the identification of the most relevant variables for each region. Furthermore, separate models were developed for each department to account for localized factors, improving overall accuracy. Exploring methods to enhance data collection, such as leveraging remote sensing technology or deploying denser sensor networks in key areas, is essential to address these limitations and improve future predictions.

Given these limitations, it is necessary to consider implementing new models that can more effectively handle the incomplete and sometimes irregular nature of the data. Models such as those based on deep neural networks or reinforcement learning techniques can be useful for dealing with large data sets with high dimensionality and possible information gaps. These models can be trained to learn complex and nonlinear patterns that might be ignored by traditional methods such as linear regression. In addition, hybrid approaches that combine different techniques, such as the use of Random Forest algorithms together with time series models, could offer a robust solution by integrating multiple data sources and providing more reliable and contextualized predictions. Thus, while data collection and management present significant challenges, the exploration of advanced, adaptive models represents a promising avenue for improving the accuracy and utility of predictive deforestation models. With an appropriate approach to data collection and the use of advanced modeling technology, it is possible to overcome these limitations and move towards more effective and sustainable conservation strategies.

We leave the deployment phase in CRISP-DM as future work, as the study is primarily research-focused. The objective is to explore and validate the model's accuracy and predictive capabilities, rather than to implement it in real-world operational systems.

9. Conclusions

The variability of deforestation data among the departments necessitated segmenting the data by region. This division enabled a more detailed and specific analysis, identifying the departments with the highest deforestation rates, such as Atlántico, Sucre, Santander, and Meta. This approach facilitates the development of more accurate predictive models adapted to each department. However, the accuracy of these models can be influenced by the variable quality of historical data, which underscores the need to improve data collection for future predictions and to ensure the applicability of models in different contexts and regions.

Linear regression models prove highly effective in predicting the annual rate of deforestation in specific departments. However, variability in data quality and socioeconomic differences between regions limit the generalizability of the results, suggesting that additional studies should be conducted before applying these models to other geographic areas or countries.

The predictive model developed showed high accuracy with metrics such as R^2, MSE, and MAE.

However, there is a possibility of bias in the data due to variability in the quality of IDEAM's historical records, which may affect the accuracy of the predictions. In addition, although the model was effective in predicting deforestation in Atlántico, Sucre, Santander, and Meta, the results may not be generalizable to other regions of Colombia or other countries due to environmental and socioeconomic differences. Caution is advised when applying these models outside the context studied.

The use of predictive models based on the CRISP-DM methodology has proven effective in predicting the deforestation rate in different regions of Colombia. Linear regression models, in particular, have demonstrated high accuracy in predicting the annual deforestation rate. This accuracy enables the early identification of critical areas and the formulation of appropriate conservation strategies.

References

[1] J. V. Solórzano, J. F. Mas, J. A. Gallardo-Cruz, Y. Gao, A. F.-M. d. Oca, Deforestation detection using a spatio-temporal deep learning approach with synthetic aperture radar and multispectral images 199 (2023) 87–101. doi:https://doi.org/10.1016/j.isprsjprs.2023.03.017.

[2] M. Leon, G. Cornejo, M. Calderón, E. González-Carrión, H. Florez, Effect of deforestation on climate change: A co-integration and causality approach with time series, Sustainability 14 (2022) 11303. doi:10.3390/su141811303.

[3] The Guardian, Deforestation in colombia falls to lowest level in 23 years (2024). URL: https://www.theguardian.com/world/article/2024/jul/10/deforestation-in-colombia-falls-to-lowest-level-in-23-years, accessed: 2024-07-11.

[4] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, A. Doulamis, A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring 34 (2023) 3299–3307. doi:10.1109/TNNLS.2022.3144791.

[5] D. C. J. Kani, S. Saudia, Analysis on the performance of machine learning models for forest fire prediction, in: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2023, pp. 1–5. doi:10.1109/ICSSIT55814.2023.10060870.

[6] J. Villalobos-Montiel, A. Aguilar-Gonzalez, L. Orona, C. Lozoya, Identifying deforested areas through convolutional neural network for drone reforesting, in: 2023 IEEE Conference on Technologies for Sustainability (SusTech), 2023, pp. 138–143. doi:10.1109/SusTech57309.2023.10129558.

[7] C. Schröer, F. Kruse, J. M. Gómez, A systematic literature review on applying CRISP-DM process model 181 (2021) 526–534. doi:10.1016/j.procs.2021.01.199.

[8] M. Hrachowitz, M. Stockinger, M. Coenders-Gerrits, R. Van Der Ent, H. Bogena, A. Lücke, C. Stumpp, Deforestation reduces the vegetation-accessible water storage in the unsaturated soil and affects catchment travel time distributions and young water fractions (2020). doi:10.5194/hess-2020-293.

[9] D. Lee, Y. Choi, MultiEarth 2022 deforestation challenge – ForestGump (2022). URL: https://arxiv.org/abs/2206.10831v1.

[10] S. Gu, The impact of increasing forest loss areas on the global temperature, and tourism industry 9 (2023) 42–55. doi:10.9734/ajraf/2023/v9i3205.

[11] R. Kumar, A. Kumar, P. Saikia, Deforestation and forests degradation impacts on the environment, in: Environmental Degradation: Challenges and Strategies for Mitigation, Springer International Publishing, 2022, pp. 19–46. doi:10.1007/978-3-030-95542-7_2.

[12] M. J. Dourojeanni, ¿es posible detener la deforestación en la amazonía peruana?, in: Desafíos y perspectivas de la situación ambiental en el Perú: en el marco de la conmemoración de los 200 años de vida republicana, Pontificia Universidad Católica del Perú, 2022, pp. 247–285. doi:10.18800/978-9972-674-30-3.013.

[13] A. Manciu, A. Rammig, A. Krause, B. R. Quesada, Impacts of land cover changes and global warming on climate in colombia during ENSO events 61 (2023) 111–129. doi:10.1007/s00382-022-06545-1.

[14] D. Mejía, M. Díaz, K. Enciso, A. Bravo, F. Florez, S. Burkart, The impact of agricultural credit on the cattle inventory and deforestation in colombia: A spatial analysis, 2022. doi:10.21203/rs.3.rs-2188032/v1.

[15] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (cnn) in vegetation remote sensing, ISPRS journal of photogrammetry and remote sensing 173 (2021) 24–49.

[16] M. H. Coelho, O. O. Bittencourt, F. Morelli, R. Santos, Método para a classificação de Áreas queimadas baseado em aprendizado de máquina automatizado 13 (2022) 029–036. doi:10.14210/cotb.v13.p029-036.

[17] A. Bommert, T. Welchowski, M. Schmid, J. Rahnenführer, Benchmark of filter methods for feature selection in high-dimensional gene expression survival data, Briefings in Bioinformatics 23 (2022) bbab354. doi:10.1093/bib/bbab354.

[18] V. Ignatenko, A. Surkov, S. Koltcov, Random forests with parametric entropy-based information gains for classification and regression problems, PeerJ Computer Science 10 (2024) e1775.

[19] J. O. Ogunleye, Predictive data analysis using linear regression and random forest, in: Data integrity and data governance, IntechOpen, 2022. doi:10.5772/intechopen.107818.