Bird Apprehension Forecasting with XGBoost

Fabrício Pereira Diniz
Paraná State University – UNESPAR
Apucarana, Paraná, Brazil
fabricio.diniz.15@estudante.unespar.edu.br

José Luis Seixas Junior
ELTE – Eötvös Loránd University,
Faculty of Informatics
Budapest, Hungary
jlseixasjr@inf.elte.hu
Paraná State University – UNESPAR
Apucarana, Paraná, Brazil
jose.junior@ies.unespar.edu.br

Guilherme Henrique de Souza Nakahata
UT – University of Tsukuba
Tsukuba, Ibaraki, Japan
GuilhermeNakahata@gmail.com
Paraná State University – UNESPAR
Apucarana, Paraná, Brazil

ABSTRACT

This article applied exploratory data analysis (EDA) and time series forecasting using extreme gradient boosting (XGBoost) to bird apprehension data from the Brazilian Institute of Environment and Renewable Natural Resources (IBAMA) covering 2010 to 2024, the dataset processed included 150,000 records, filtered to focus on significant patterns across two distinct periods: a COVID-19 pre-pandemic period (2010-2020) and a period that includes the COVID-19 pandemic (2014-2024). This analysis identified shifts in apprehension patterns with an overall decrease of 69.17% in bird apprehensions during the pandemic. The XGBoost model demonstrated satisfactory root mean squared error (RMSE) on most of the species except on *Zenaida auriculata* and *Zenaida auriculata noronha* where the RMSE values were 37.73 and 47.80, respectively, for the (2010–2020) period, while for the (2014–2024) period, the RMSE values were 42.93 and 48.76, which are higher values when compared to other results. The most apprehended bird species pre-pandemic was *Sicalis flaveola* with 163, 437 apprehensions, while *Zenaida auriculata* and *Zenaida auriculata noronha* remained highly apprehended throughout both periods. The northeast regions of Brazil, with states like *Rio Grande do Norte*, *Ceará* and *Paraíba*, showed the highest apprehension rates. Overall, the model achieved satisfactory performance in understanding the pattern of apprehension numbers in the tested periods, with significantly low RMSE values for the ten most apprehended species.

Keywords

Bird Apprehension, IBAMA, Time Series Forecasting, XGBoost

1 INTRODUCTION

Brazilian territorial extension occupies 47.3% of South America [1], being home to approximately 57% of the bird species recorded in South America, more than 10% of these species are endemic to the country, which brings a great responsibility to the country to preserve these species, highlighting the attention needed on investment in conservation efforts [2].

Brazil serves as a key region for understanding the dynamics of species transactions and smuggling [3], however, due to the fact that it is a trafficking route for these animals, the Brazilian police are able to make a high number of apprehensions that are recorded by the Ministry of the Environment. These apprehension statistics reveal possible impacts on their wild populations as well as on the ecosystems in which they are part, being integral part of the biological diversity.

Based on these records, it becomes possible to perform an exploratory analysis to better visualize and understand possible trends and train forecasting models to help us understand patterns and help the police and ministry make decisions about possible investments to be made in this important preservation task.

To carry out this work, Exploratory Data Analysis (EDA) was first applied to explore and understand the patterns, trends, and locations with higher incidence of occurrences. EDA helped identify the periods when these events occurred more frequently. After this preliminary analysis, Extreme Gradient Boosting (XGBoost) was used to model and make predictions based on the features identified during the data exploration phase, this research offers a comparative analysis between a COVID-19 pre-pandemic period (2010-2020) and a period that includes the COVID-19 pandemic (2014-2024).

Thus, the objective of this work is to forecast the ten most apprehended birds in the last decade in Brazilian territory, the locations and their possible temporal characteristics, by using data recorded since February 22, 1989, the date of creation of the Brazilian Institute of Environment and Renewable Natural Resources[1] (IBAMA).

The article is organized as follows: Section 2 describes the related works; Section 3 shows the proposed method; Section 4 describes the experimental results and discusses the results; Finally, the conclusions are described in Section 5.

2 RELATED WORKS

The use of XGBoost for time series forecasting is well-documented across various fields [4–7], such as in the financial market [8, 9], predicting cases of COVID-19 [10, 11] and in predicting type 1 and 2 diabetes [12, 13]. These articles obtained results explored in the context of this work, and they were incorporated as the basis for this study.

An environmental related study using the Extreme Gradient Boosting has explored predicting the primary lifestyle of birds based on morphological traits. Although forecasting was not applied, XGBoost was successfully applied to achieve results with the AVONET dataset [14]. These papers, linked to the fact that tree algorithms still achieve good results even when compared to deep models [15], give us the perception that this algorithm is quite suitable for trying to solve the problem at hand.

Alencar and Fonseca [16] presented a study on the ornithofauna apprehended between 2013 and 2022. Their work, published in *Revista Agrogeoambiental*, utilized the same database but covered a different time period. While Alencar and Fonseca aimed to evaluate

[1] Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis

the illegal trade of wild birds in Brazil by analyzing, identifying, and quantifying birds apprehended by IBAMA and Federal Highway Police (PRF[2]) in Brazilian territory.

3 PROPOSED METHOD

IBAMA, as an autarchy linked to the Brazilian Ministry of the Environment and Climate Change[3], is responsible for implementing national environmental policies and developing preservation and conservation actions, exercising control and inspection of natural resources. It is, therefore, responsible for managing data on animal apprehension made by the Brazilian federal police.

Therefore, when species are apprehended, the data is recorded and made publicly available by IBAMA[4], based on these records, an exploratory analysis becomes possible to better visualize and understand possible patterns, making it easier to understand how the forecasting model might perform [17].

Given its tabular nature and timeline structure, as seen in the related works, this data refinement process, through EDA, made it possible to carry out a time series forecasting algorithm. When compared to other gradient boosting implementations, the XGBoost algorithm is faster and excels in handling tabular datasets in regression predictive modeling problems, while also supporting continued training [18].

These methods combined resulted in accurate predictions with minimal training time, as the time period presented in this article includes a pandemic period and this period caused large distortions in the numbers, the chosen time frame was divided in two segments to allow for a broad analysis. One starting at the last data and ten years backwards (2014-2024) and another starting before the pandemic and ten years backwards (2010-2020).

3.1 Database

The initial dataset[5] was obtained from IBAMA's open data. It is a comma-separated values file with 35 columns and 150.000 rows. As this is a Brazilian dataset, column names and table data appear in Portuguese. The first step was to separate the file into the analyzed periods based on "DAT_TAD" column, which showed the date the term was drawn up. The data was separated in two periods: 2014 to 2024, including the pandemic period, and 2010 to 2020, representing a pre-pandemic period.

After separating the data based on the "SIT_CANCELADO" column, which indicates whether the formally issued term was canceled, the rows were removed where the term had been canceled or were inconsistent, meaning they contained invalid values. As the data table, even after filtering and corrections, still had several missing values, an additional IBAMA's dataset was used. The "especime_apreendido" table contains the column "SEQ_TAD", a column that represents the key that identifies the term of inspection, also present in the "termo_apreensao" table. Using these identifiers, two spreadsheets were merged through the list of identifications.

Two columns of scientific name (*Nome Científico*) and popular name (*Nome Popular*) were used to identify the bird species. There were gaps where IBAMA did not provide information. Since the rest of the line was consistent, an association was made between the two columns regarding names. Specifically, using the information provided by a row with both fields filled in, a search was conducted to fill in the scientific name for all rows with the same popular name.

The group column (*GRUPO*) was extremely important for filtering only bird-related data, where the rows labeled as "Bird" (*Ave*) were used to fill in any gaps. The lines corresponding to the group 'bird' were selected, and the lines where the description (*DESCRICAO_PRODUTO*) was 'parts' (*Partes*) were removed to ensure more accurate data by excluding bird parts to be taken into account during the experiments, there was no detailed information regarding these parts, such as which body pieces they represented or whether they came from a single specimen or multiple ones.

The quantity was standardized by retaining lines where it was indicated an unit in column "DES_UNIDADE_MEDIDA", while disregarding cubic meters or kilograms. This resulted in the creation of a new comma-separated values file containing precise quantities and key information as illustrated in the flowchart in Figure 1.

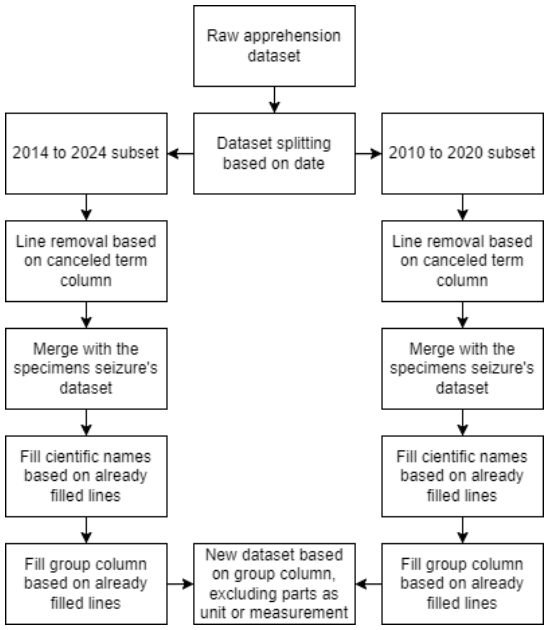

Figure 1: Dataset preparation flowchart

3.2 Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble model that leverage on weak models, such as decision trees, which are good generalizers so that the combined actions of these regressors constitute a good final model [19]. Based on the work of Friedman [20], the ensemble of trees are iteratively adjusted to minimize an objective function. This objective function incorporates regularization terms that control the model's complexity, promoting simpler and more robust solutions.

[2]Polícia Rodoviária Federal
[3]Ministério do Meio ambiente e Mudança Climática
[4]https://dadosabertos.ibama.gov.br/dataset/fiscalizacao-termo-de-apreensao/resource/3364a902-067a-469a-bd9a-b865d234eed9
[5]https://dadosabertos.ibama.gov.br/dataset/fiscalizacao-termo-de-apreensao/resource/081627a0-d158-44a8-88b0-600929e93526

The XGBoost framework, created by Chen and Guestrin [21] and provided by the Distributed (Deep) Machine Learning Community (DMLC) team[6], was adopted using a learning rate of 0.01 and 1000 estimators, along with 50 rounds for early stopping, preventing overfitting while ensuring sufficient model training, all other hyper parameter values were kept as the default settings, the parameters of the XGBoost were adjusted in order to perform a regression task on tabular data for each species.

The model continues to be a state-of-the-art approach, Luo et al. [22] integrating Convolutional Neural Network-Bidirectional Long Short-Term Neural Network-Attention Mechanism (CNN-BiLSTM-Attention) with XGBoost, stacking the models to forecast short-term load data. Kazemi et al. [23] proposed a novel hybrid methodology integrating XGBoost with Harris Hawk optimization and dragonfly algorithm to enhance penetration rate prediction in rotary drilling. Novel approaches do not necessarily need to directly modify the model; instead, they suggest implementations alongside other techniques to enhance performance in specific problems if needed.

XGBoost efficiently handle high-dimensional and large-volume datasets, which are often encountered in practical machine learning challenges. Its flexibility in parameter tuning allows the model to be adapted to different types of tasks, including regression and classification. Experiments reported in the study by Chen and Guestrin [21] demonstrated that the system achieves state-of-the-art results on standard benchmarks, solidifying its position as a reference tool in the field.

3.3 Metrics

The model evaluation was based on the Root Mean Square Error (RMSE), a simple and common metric used to measure the distance of the predicted and real values [24],

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i is the sample observed value and $\hat{y}_i$ the predicted value by the model for sample i. When evaluating the model using

$$(y_i - \hat{y}_i)^2$$

the exponentiation amplifies larger errors since an error of greater magnitude will have a quadratic impact on the total sum, while smaller errors will have a relatively minor impact. As there are many spikes in the data that need to be considered in the analysis, RMSE was utilized to penalize large errors, making the evaluation of the model more sensitive to extreme variations.

A high value for RMSE indicates a significant discrepancy or error between the observed values and the predicted values, therefore, a substantial difference between the two sequences or datasets being analyzed. A low value suggests that the difference is small, indicating good accuracy or a closer match between the sequences, while a value of zero, in turn, represents a perfect match with no discrepancy between the two sequences.

4 EXPERIMENTAL RESULTS AND DISCUSSION

It should be noted that the objective of the work is to analyze it over ten years, however, starting from 2024 and going back ten years, the period incorporates the pandemic, which would cause drastic changes in the final period of the dataset. In other words, the intention was to do this analysis 2024 and ten years ago, but it would result in a portion of data that models or EDA would have a lot of trouble analyzing.

Then, another "clean" dataset was created with pre-pandemic data, therefore, the initial dataset was divided into two distinct ten-year periods: a pre-pandemic period (2010-2020) and a period that includes the pandemic (2014-2024).

This separation was made due to differences observed during the EDA process in the processed dataset from 2014 to 2024, as can be seen a significant difference on the apprehension values in the data, where the mean was $\overline{x} = 13793.00$ and the median $\tilde{x} = 12927.00$ in the pre-pandemic years (2014 - 2020), and $\tilde{x} = 12927.00$ and $\tilde{x} = 3282.50$ in the period after the start of the pandemic (2020 - 2024), highlighting the problem os containing a pandemic portion.

Given this unexpected event, the model's performance was analyzed across these two time periods. Therefore, both the analysis carried out only within the period and the analysis excluding the period would be deficient, as the pre-pandemic numbers no longer represent the reality of apprehensions, just as the pandemic period no longer represents the global situation at the moment.

The numbers suggest a substantial decrease in the number of apprehensions after the start of the pandemic, which could be attributed to several factors such as changes in behavior, restrictions, fewer outliers or changes in law enforcement priorities during the pandemic period. The two periods distributions can be seen in Figure 2.

In the two new final processed datasets, it was possible to identify the number of apprehensions per species and their respective locations through the columns longitude and latitude of terms, showing that most apprehensions were carried out in the northern part of Brazil. The number of most apprehended birds in the different periods analyzed are listed in Tables 1 and 2.

Table 1: Most apprehended birds in the COVID-19 pre-pandemic period (2010 - 2020)

Scientific name	Quantity
Sicalis flaveola	163,437
Oryzoborus maximiliani	48,397
Zenaida auriculata	31,400
Zenaida auriculata noronha	27,394
Oryzoborus angolensis	24,754
Gallus domesticus	23,286
Sporophila caerulescens	10,085
Paroaria dominicana	9,724
Saltator similis	9,221
Sporophila albogularis	6,903

The total quantity of birds apprehended among the ten most incidents fell from 354, 601 to 109, 314, this represents ≈ 69.17% fewer apprehensions. Species like *Oryzoborus maximiliani* and *Gallus domesticus* no longer appear among the most apprehended, while *Sporophila nigricollis* and *Sporophila lineola* were among the most recorded.

[6]https://xgboost.readthedocs.io

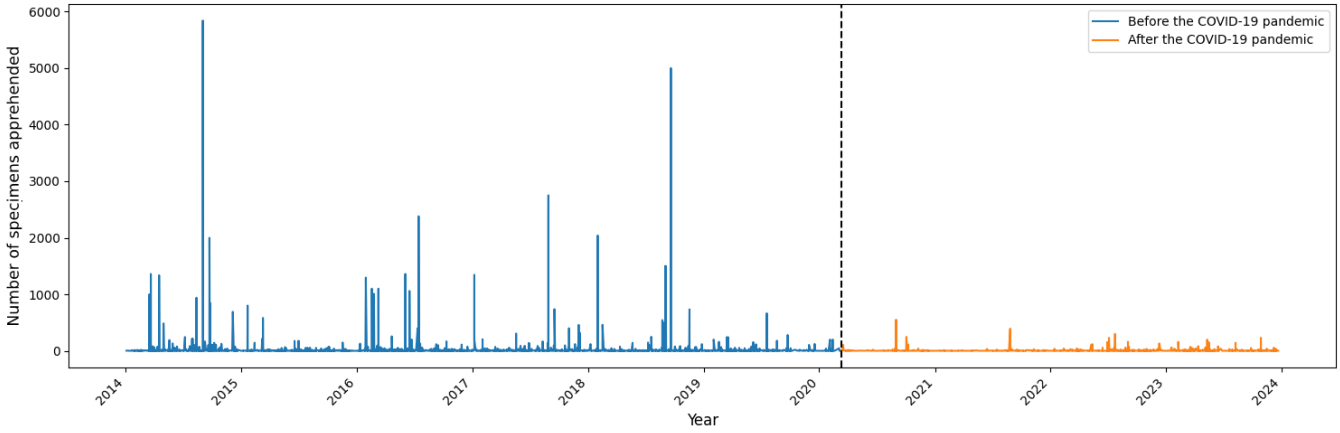

Figure 2: Number of species apprehended from 2014 to 2024 in the processed dataset

Table 2: Most apprehended birds including the COVID-19 pandemic period (2014 - 2024)

Scientific name	Quantity
Zenaida auriculata	27,729
Zenaida auriculata noronha	25,730
Sicalis flaveola	20,745
Paroaria dominicana	7,333
Saltator similis	6,316
Sporophila caerulescens	6,279
Sporophila nigricollis	4,947
Oryzoborus angolensis	3,883
Sporophila albogularis	3,620
Sporophila lineola	2,732

However, during the COVID-19 pandemic, apprehensions of *Zenaida auriculata*'s remained at levels similar to the pre-pandemic ones, while other *passeriformes* species showed significant declines in apprehension numbers. As observed in the previously mentioned work by Alencar and Fonseca [16], this species is the only columbiforme species on the list, reaffirming the point that this species is intended for consumption in Brazil's northeastern region.

Table 3: Number of apprehensions in the five states with the most apprehensions in the COVID-19 pre-pandemic period (2010 - 2020)

State	Total apprehended	Most apprehended species	Quantity	Percentage of State Total
RN	197,209	*Sicalis flaveola*	134,174	68.04%
RS	46,794	*Oryzoborus maximiliani*	46,041	98.39%
CE	24,718	*Zenaida auriculata*	8,423	34.08%
PB	15,904	*Zenaida auriculata*	4,419	27.79%
PA	15,094	*Oryzoborus angolensis*	12,831	85.01%

By individually analyzing the states and most apprehended species in each of them, it is observed that *Rio Grande do Norte* (RN), *Ceará* (CE) and *Paraíba* (PB) remain among the five states with the highest number of apprehensions. However, while *Rio Grande do Sul* (RS) and *Pará* (PA) were among the states with the most apprehensions before the pandemic, this ranking is now occupied by *Bahia* and *Pernambuco* during the pandemic period. The number of apprehensions in the five states with the most apprehensions, considering and not considering the COVID-19 pandemic period, can be seen in Tables 3 and 4, where the column "Percentage of State Total" indicates the percentage contribution of the most apprehended species to the total apprehensions in each state.

Table 4: Number of apprehensions in the five states with the most apprehensions in the period that includes the COVID-19 pandemic (2014 - 2024)

State	Total apprehended	Most apprehended species	Quantity	Percentage of State Total
RN	43,772	*Zenaida auriculata noronha*	21,888	50.00%
CE	13,521	*Zenaida auriculata*	4,147	30.68%
PB	9,486	*Zenaida auriculata*	4,272	45.01%
BA	8,001	*Sicalis flaveola*	4,488	56.09%
PE	5,985	*Sicalis flaveola*	2,041	34.10%

A preprocessing step was carried out by applying the Inter Quartile Range (IQR) method due to its simplicity and effectiveness to remove outliers [25]. When applying XGBoost, each species were selected individually from the dataset, and separated 2/3 for training and 1/3 for testing using a block-wise split, which partitions data into sequential, non-randomized subsets to maintain chronological integrity and prevent overestimation of model performance. The features used include time, day of the week, quarter, month, year and day of the year.

Hyperparameters were arranged as follows: the base score was set to 0.5, the booster type was specified as 'gbtree', the number of estimators was set to 1000, the early stopping rounds were adjusted to 50, the objective function was defined as 'reg:squarederror', the maximum depth was set to 3, and the learning rate was configured to 0.01. The model was evaluated using Root Mean Square Error

(RMSE) where the highest values were from *Zenaida auriculata noronha* and *Zenaida auriculata*, as show in Table 5.

Table 5: RMSE values for different species over two periods

Species	2010 to 2020 data's RMSE	2014 to 2024 data's RMSE
Gallus domesticus	7.01	—
Oryzoborus angolensis	1.33	1.21
Oryzoborus maximiliani	1.41	—
Paroaria dominicana	1.30	1.15
Saltator similis	1.52	1.82
Sicalis flaveola	2.85	2.08
Sporophila albogularis	1.22	0.68
Sporophila caerulescens	1.37	1.34
Sporophila lineola	—	0.75
Sporophila nigricollis	—	1.38
Zenaida auriculata noronha	47.80	48.76
Zenaida auriculata	37.73	42.93

Elements with no value (symbolized with —) in the Table 5 refer to species that were not among the ten most apprehended during the period. The fact that they are not on the list of the ten may result in there not being enough records or patterns to be detected by the model, and this is the reason why the analyses were not performed with all species but only on the ten most relevant during the period.

The presented values indicate the accuracy of XGBoost for different species in different periods, being measured by the RMSE, a accuracy assessment metric that measures the difference between the values predicted by the model and the actual observed values. No data normalization was applied to maintain the integrity of the values, as the model works well with regressions, this factor is not particularly significant, it should only be taken into account when analyzing the results.

The analysis is not significantly impacted by any normalization because the classes with the highest values are highly discrepant from the rest, which is not the case with the number of occurrences, therefore, even though *Zenaida auriculata* and *Zenaida auriculata noronha* resulted in high RMSE values, this suggests that the model had difficulty accurately predicting apprehensions of these species. While *Sporophila albogularis* had the lowest observed values, indicating more accurate predictions over the analyzed time.

Zenaida auriculata noronha which is most often captured for consumption in northeastern Brazil, maintained similar values across both analyzed periods, having the worst performances in the RMSE. This happens because when captured they were dead to serve as food, that is, in greater quantities creating spikes in the data making it more difficult to the model to have a good performance. This species might have such a distinct characteristic that it may require a separate study, or a specific model adjustment to be able to handle.

Sicalis flaveola is another species worth mentioning, as it had a greater number of occurrences pre-pandemic, and still obtained a low RMSE value. Without normalization, this indicate that the model managed to capture the pattern of this species well.

5 CONCLUSION

Experimental results showed that although the pandemic period was associated with a decrease in apprehensions among the most frequently apprehended birds, the RMSE remained very close to the results of the previous period indicating consistent performance on both time intervals.

Species such as *Oryzoborus maximiliani* and *Gallus domesticus*, which were not among the most apprehended in the pre-pandemic period, did not appear in the most recent data. However, *Sporophila nigricollis* and *Sporophila lineola* appeared among the most recorded species, this could reflect shifts in illegal bird trade dynamics, potentially requiring targeted interventions.

The analysis by states revealed that Rio Grande do Norte, Ceará and Paraíba remained among the states with the highest number of apprehensions, Rio Grande do Sul and Pará were replaced by Bahia and Pernambuco during the pandemic period.

Despite the challenges posed by spikes in certain species and regional variations, the satisfactory RMSE values demonstrate the feasibility of using IBAMA's open data for predictive modeling, an approach that had not previously been explored with forecasting techniques. The species *Zenaida auriculata noronha* presented particular challenges and exhibited poor RMSE performance. The main hypothesis for this is that most of the captures involve dead species intended for consumption in northeastern Brazil, which creates several data spikes, making it difficult for the model to perform accurately. This suggests that *Zenaida auriculata noronha* may require a separate study to better capture its unique pattern.

These insights can be instrumental for IBAMA in identifying emerging trends in wildlife trafficking, prioritizing high-risk regions, and focusing conservation efforts on species under increasing pressure.

Alternatives for future works include implementing and evaluating other forecasting models, as well as applying spatiotemporal clustering algorithms. The results could also be used to guide the creation of an image dataset featuring the most apprehended birds, making it possible to train classification models. Such models could automate the identification of species in field inspections, helping IBAMA collect more accurate data, which could be used later to combat wildlife trafficking more effectively.

6 ACKNOWLEDGMENT

We would like to express our sincere gratitude to the State University of Paraná (UNESPAR) for their support throughout this research.

REFERENCES

[1] Antônio Herman Vasconcellos Benjamin. Introdução ao direito ambiental brasileiro. *Revista de direito ambiental*, 4(14):48–82, 1999. doi: 10.22456/2317-8558.49540.

[2] Miguel Angelo Marini and Frederico Innecco Garcia. Bird conservation in brazil. *Conservation Biology*, 19(3):665–671, 2005. doi: 10.1111/j.1523-1739.2005.00706.x.

[3] Rômulo Romeu Nóbrega Alves, José Ribamar de Farias Lima, and Helder Farias Pereira de Araujo. The live bird trade in brazil and its conservation implications: an overview. *Bird Conservation International*, 23(1):53–65, 2013. doi: 10.1017/S095927091200010X.

[4] Shobhit K. Patel, Jaymit Surve, Vijay Katkar, Juveriya Parmar, Fahad Ahmed Al-Zahrani, Kawsar Ahmed, and Francis Minhthang Bui. Encoding and tuning of thz metasurface-based refractive index sensor with behavior prediction using xgboost regressor. *IEEE Access*, 10:24797–24814, 2022. doi: 10.1109/ACCESS.2022.3154386.

[5] Dennis A-L Mariadass, Ervin Gubin Moung, Maisarah Mohd Sufian, and Ali Farzamnia. Extreme gradient boosting (xgboost) regressor and shapley additive explanation for crop yield prediction in agriculture. In *12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 219–224, 2022. doi: 10.1109/ICCKE57176.2022.9960069.

[6] Selvaprabu Jeganathan, Arun Raj Lakshminarayanan, Nandhakumar Ramachandran, and Godwin Brown Tunze. Predicting Academic Performance of Immigrant Students Using XGBoost Regressor. *International Journal of Information Technology and Web Engineering (IJITWE)*, 17(1):1–19, January 2022. doi: 10.4018/IJITWE.304052.

[7] Xinmeng Zhang, Chao Yan, Cheng Gao, Bradley A. Malin, and You Chen. Predicting Missing Values in Medical Data Via XGBoost Regression. *Journal of Healthcare Informatics Research*, 4(1):383–394, August 2020. doi: 10.1007/s41666-020-00077-1.

[8] Yan Wang and Yuankai Guo. Forecasting method of stock market volatility in time series data based on mixed model of arima and xgboost. *China Communications*, 17(3):205–221, 2020. doi: 10.23919/JCC.2020.03.017.

[9] Pham Hoang Vuong, Trinh Tan Dat, Tieu Khoi Mai, Pham Hoang Uyen, and Pham The Bao. Stock-price forecasting based on xgboost and lstm. *Computer Systems Science and Engineering*, 40(1):237–246, 2022. doi: 10.32604/csse.2022.017685.

[10] Junling Luo, Zhongliang Zhang, Yao Fu, and Feng Rao. Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms. *Results in Physics*, 27:104462, 2021. ISSN 2211-3797. doi: 10.1016/j.rinp.2021.104462.

[11] Md. Siddikur Rahman, Arman Hossain Chowdhury, and Miftahuzzannat Amrin. Accuracy comparison of arima and xgboost forecasting models in predicting the incidence of covid-19 in bangladesh. *PLOS Global Public Health*, 2(5):1–13, May 2022. doi: 10.1371/journal.pgph.0000495.

[12] Cooper Midroni, Peter J Leimbigler, Gaurav Baruah, Maheedhar Kolla, Alfred J Whitehead, and Yan Fossat. Predicting glycemia in type 1 diabetes patients: experiments with xgboost. 60(90):120, 2018.

[13] Liyang Wang, Xiaoya Wang, Angxuan Chen, Xian Jin, and Huilian Che. Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. volume 8, page 247, 2020. doi: 10.3390/healthcare8030247.

[14] Luis Javier Madrigal-Roca. Assessing the predictive value of morphological traits on primary lifestyle of birds through the extreme gradient boosting algorithm. *PLOS ONE*, 19(1):1–18, Jan 2024. doi: 10.1371/journal.pone.0295182.

[15] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. doi: dl.acm.org/doi/10.5555/3600270.3600307.

[16] Leonardo Alencar and Rogério Fonseca. Ornithofauna seized by enforcement agencies from 2013 to 2022. *Revista Agrogeoambiental*, 16(unico):e20241829–e20241829, 2024. doi: 10.18406/2316-1817v16nunico20241829.

[17] Chris Chatfield. Exploratory data analysis. *European journal of operational research*, 23(1):5–13, 1986. doi: 10.1016/0377-2217(86)90209-2.

[18] Jason Brownlee. *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016.

[19] Majid Niazkar, Andrea Menapace, Bruno Brentan, Reza Piraei, David Jimenez, Pranav Dhawan, and Maurizio Righetti. Applications of xgboost in water resources engineering: A systematic literature review (dec 2018–may 2023). *Environmental Modelling & Software*, 174:105971, 2024. ISSN 1364-8152. doi: 10.1016/j.envsoft.2024.105971.

[20] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451.

[21] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

[22] Shucheng Luo, Baoshi Wang, Qingzhong Gao, Yibao Wang, and Xinfu Pang. Stacking integration algorithm based on cnn-bilstm-attention with xgboost for short-term electricity load forecasting. *Energy Reports*, 12:2676–2689, 2024. ISSN 2352-4847. doi: 10.1016/j.egyr.2024.08.078.

[23] Mohammad Mirzehi Kalate Kazemi, Zohre Nabavi, and Danial Jahed Armaghani. A novel hybrid xgboost methodology in predicting penetration rate of rotary based on rock-mass and material properties. *Arabian Journal for Science and Engineering*, 49(4):5225–5241, 2024.

[24] Timothy O. Hodson. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 15:1–10, 2022. doi: 10.5194/gmd-15-5481-2022.

[25] Giulio Barbato, EM Barini, Gianfranco Genta, and Raffaello Levi. Features and performance of some outlier detection methods. *Journal of Applied Statistics*, 38 (10):2133–2149, 2011. doi: 10.1080/02664763.2010.545119.