Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4230

Enhancing the Performance of Machine Learning Classifiers through Data Cleaning with Ensemble Confident Learning

Renato Okabayashi Miyaji [**Escola Politécnica da Universidade de São Paulo** | ***re.miyaji@usp.br***]
Felipe Valencia de Almeida [**Escola Politécnica da Universidade de São Paulo** | ***felipe.valencia.almeida@usp.br***]
Pedro Luiz Pizzigatti Corrêa [**Escola Politécnica da Universidade de São Paulo** | ***pedro.correa@usp.br***]

Escola Politécnica, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 380 - Butantã, São Paulo - SP, 05508-010, Brazil.

Received: 9 March 2024 • **Published:** 20 January 2025

Abstract Model-centric techniques, such as hyper parameter optimization and regularization, are commonly used in the literature to enhance the performance of Machine Learning Classifiers. However, when dealing with noisy data, Data-Centric approaches show promising potential. Thus, in this paper a new method is proposed: the Ensemble Confident Learning (ECL), which enhances the Confident Learning technique with the use of multiple learners to improve the selection of instances with biased labels. This method was applied for a case study of Species Distribution Modeling in the Amazon using Classifiers to estimate the probability of species occurrence based on environmental conditions. Compared to Confident Learning, ECL showed an improvement of 20% in Recall and 3.5% in ROC-AUC for Logistic Regression.

Keywords: Data Cleaning, Machine Learning, Confident Learning, Species Distribution Modeling

1 Introduction

To enhance the performance of Machine Learning Classifiers, techniques that aim at improving their training process or optimizing their hyper parameters are commonly applied [James *et al*., 2013]. Examples of such methods are Lasso [Tibshirani, 1996] and Ridge [Hoerl and Kennard, 1970] regularizations in the case of Linear Regression.

In the literature, other techniques have been developed for this purpose. However, they seek to systematically modify the training dataset to improve the performance of Machine Learning Classifiers. This approach is known as Data-Centric Artificial Intelligence (DCAI) and differs from the other common methods that aim at improving models: Model-Centric Artificial Intelligence (MCAI) [Hamid, 2022].

The Data-Centric approach (DCAI) is useful in various contexts, especially when the dataset used for modeling is limited – when collecting more data is complex or impossible – or when there are uncertainties regarding the collected data [Hamid, 2022]. Thus, there are two classes of techniques: those that use algorithms to gain a better understanding of the data and utilize this information to enhance the training process of models, such as Curriculum Learning [Bengio *et al*., 2009]; and algorithms that modify the training data to improve the performance of Machine Learning Classifiers [Northcutt *et al*., 2021b].

One such technique is proposed by Northcutt *et al*. [2021b], called Confident Learning (CL). As a DCAI method, it focuses on assessing the quality of labels of the target variable by characterizing and identifying errors. This is done based on the principles of noisy data pruning, with the definition of probabilistic boundaries to estimate uncertainties and then ordering uncertain instances and training the model on a more confident dataset.

Confident Learning techniques have been successfully used in the literature. Northcutt *et al*. [2021a] applied to the 10 most common datasets in Data Science used for Computer Vision, Natural Language Processing, among others. By using CL, it was possible to identify an average of 3.3% of the total instances with noisy labels, which when treated or removed, resulted in better performance of the Machine Learning Classifiers used for prediction.

However, in the literature it is stated that one of the limitations of Confident Learning is that the method is based on the process of determining the self-confidence of the model to classify instances according to its confidence in their labels. This fact can lead to biased results when the model does not have a good predictive ability. In this scenario, a low confidence of the model for a class can lead to a considerable amount of noisy instances being classified as correct and, therefore, not treated [Zhang *et al*., 2023].

Thus, in the work of Zhang *et al*. [2023] a more robust method is proposed based on Confident Learning. The new method brings two new contributions in comparison with the original one: the selection of instances with noisy labels is done by two different models that are trained in different datasets, as the Co-teaching [Han *et al*., 2018] method proposes; and the predictions and thresholds of both models are used to compute a confidence interval to determine if an instance has a noisy label or not. Li *et al*. [2023] proposed another method based on Confident Learning, called Decoupled Confident Learning (DeCoLe). It differs from the method proposed by Northcutt *et al*. [2021b], since it does not hold the hypothesis of class conditional noise and, therefore, is more general than the original one.

In this context, this extended paper from the one previously published by Miyaji *et al.* [2023] aimed to propose a new Data-Centric Artificial Intelligence (DCAI) method based on Confident Learning - the Ensemble Confident Learning (ECL) - and compare it with the original one in a case study with the objective of cleaning a dataset and addressing uncertainties within it to enhance the performance of Machine Learning Classifiers.

This paper is organized into five main sections. Section 1 provides an introduction to Confident Learning based methods and the case study. Section 2 discusses the related work and literature review. Section 3 details the methodology of Confident Learning and Ensemble Confident Learning, as well as the case study. Section 4 presents the results of the case study and discussions. Finally, the Conclusion section summarizes the key findings, discusses the limitations of the study, and suggests directions for future research.

2 Related Works

Standardizing inconsistent data is one of the main challenges in databases and related applications, such as analytics and predictive modeling. Therefore, different techniques for automatically cleaning datasets and handling their uncertainties to enable the application of Machine Learning Classifiers have been developed in the literature. In the works of Elcan [2001] and Forman [2005], techniques aiming to estimate false positive and false negative rates for Binary Classification tasks, such as Cost-Sensitive Learning, were proposed.

In Elcan and Noto [2008], a new technique for a more robust Binary Classification was proposed. It can be applied to problems in which only data regarding the positive class is available, along with unlabeled instances for the target variable. This is achieved by introducing the concept of classification threshold to identify uncertain instances in the dataset. However, its main limitation is the requirement for a fully reliable positive class data.

Lipton *et al.* [2018] propose the Black Box Shift Estimation (BBSE) technique to identify instances with inverted or noisy labels, using Confusion Matrices and Cross-Validation processes. In the work of Huang *et al.* [2019], the effectiveness of identifying uncertain instances, treating them, and then training a Machine Learning model on the treated dataset with a significant performance gain is demonstrated.

The Confident Learning (CL) technique, proposed by Northcutt *et al.* [2021b], aimed to incorporate the main contributions of the previously mentioned authors to obtain a more robust and generalized method capable of handling uncertain and/or unlabeled instances, identifying and treating them, and being applicable to any kind of Machine Learning Classifier, as well as multi-class classification problems.

In the works of Zhang *et al.* [2023] and Li *et al.* [2023], Confident Learning based methods were proposed to address the limitations of the original method. The method proposed by Zhang *et al.* [2023] improves the robustness of the process of selecting instances with noisy labels using two models, as it happens in Co-teaching [Han *et al.*, 2018], and calculating confidence intervals. The Decoupled Confident Learning (DeCoLe) method also makes the selection of noisy labels more robust by training one model for each group and defining upper and lower bounds [Li *et al.*, 2023].

Differently from the methods proposed by Zhang *et al.* [2023] and Li *et al.* [2023], Ensemble Confident Learning aims to address the limitation of the original method of the high dependency on the self-confidence by considering multiple learners to determine whether an instance has a noisy label or not. In order to achieve that, different models are trained in datasets with different frequencies of samples belonging to the minority and majority classes, inspired by techniques developed to mitigate label noise and class imbalance during the training process of Neural Networks, such as Co-teaching [Han *et al.*, 2018] and Balanced Mini-Batch Training [Shimizu *et al.*, 2018].

The Co-teaching method was developed to enhance the learning process of Neural Networks, when there are noisy labels in the dataset. This is achieved through the training of two Networks simultaneously. The instances with noisy labels can be identified through their losses and both Networks communicate with each other during the training process [Han *et al.*, 2018]. The Balanced Mini-Batch Training technique aims at enabling the training process of Neural Networks for Classification tasks, when there is an imbalanced dataset. Thus, during the training process of the Network, resampling techniques are used to balance each mini-batch of the data [Shimizu *et al.*, 2018].

In ECL, concepts of both methods were used to enhance the Confident Learning method. The different proportions between samples of the minority and majority classes are generated with Imbalanced Learning techniques, such as Synthetic Minority Oversampling Technique (SMOTE) [The Imbalanced-learn Developers, 2024]. To classify a label as noisy, a voting between the learners is performed, as it happens in Ensemble methods [James *et al.*, 2013].

The Ensemble Confident Learning (ECL) method brings contributions to the literature of learning with label noise, such as: it addresses the risk of biased outputs resulting from the Meta-Learning process of Confident Learning with the use of multiple learners (Ensemble); it is also able to treat Class Imbalance tasks with Imbalanced Learning techniques; and it is model-agnostic.

Table 1 summarizes the difference between the Confident Learning based methods. All of them are based on the Meta-Learning process for Noisy Data Pruning. The methods Decoupled Confident Learning (DeCoLe) [Li *et al.*, 2023], Fairness Confident Learning [Zhang *et al.*, 2023] and Ensemble Confident Learning use multiple models to improve the confidence in the process of identifying instances with noisy labels. The only method that is capable of enabling the learning when there is a class imbalance is Ensemble Confident Learning, since it uses Imbalanced Learning techniques.

To compare the new method with the original one and model-centric techniques, a case study regarding a Species Distribution Modeling (SDM) experiment was conducted, in which Classifiers are used to estimate the probability of species occurrence based on environmental variables.

Species Distribution Models (SDMs) are widely used in Ecology to quantitatively assess the ecological niche of the

Table 1. Comparison between Confident Learning based methods: Confident Learning (CL), Fairness Confident Learning (FCL), Decoupled Confident Learning (DeCoLe) and Ensemble Confident Learning (ECL)

Method	Noisy Data Pruning	Multiple Models	Imbalanced Data
CL	X		
FCL	X	X	
DeCoLe	X	X	
ECL	X	X	X

analyzed species, which is the range of values of environmental variables that make an habitat suitable for the occurrence of the species under study [Hutchinson, 1991]. In recent decades, there has been a significant advancement in the field of Machine Learning, with the development of models capable of impressive performances. Therefore, these models have been increasingly used for Species Distribution Modeling problems [Hegel *et al*., 2010].

However, Machine Learning models show greater potential for tasks, in which a large and reliable dataset is available for modeling. Specifically for Species Distribution Modeling, this is not the case, as occurrence data of species are typically used. These data are often provided in the form of presence-only data. The most common datasets contain only instances of species presence - equivalent to the positive class for the Classification problem - while the assertion of species absence - equivalent to the negative class - is a process that introduces uncertainties into the data. A common practice is to incorporate pseudo-negative samples (or Pseudo-absences) to represent the negative class of the dataset [Beery *et al*., 2021].

However, using biased data in the SDM process can lead to incorrect results [Martin *et al*., 2005]. In this scenario, by enabling the treatment of uncertainties through data cleaning, the Ensemble Confident Learning (ECL) and Confident Learning (CL) techniques demonstrate great potential to enhance the use of Machine Learning Classifiers for Species Distribution Modeling.

Specifically for the task of Species Distribution Modeling, several approaches have been adopted in the literature for handling uncertainties. One of them is through the Bayesian approach, which enables the incorporation of expert knowledge or previous studies into the models using the Bayes Theorem. In Bayesian models, prior and likelihood probability distributions can be provided for their parameters, from which the posterior probability distribution is determined. The Bayesian Logistic Regression classifier, evaluated by Miyaji and Corrêa [2021], Di Lorenzo *et al*. [2011], and Golini [2011], showed great potential for handling uncertainties. However, the main limitation of this approach is the requirement for expert knowledge or previous studies for its application, and it is not applicable to all kinds of Classifiers.

Another approach is to represent negative classes in the dataset through pseudo-negative samples [Beery *et al*., 2021]. Preferably, these should be selected based on expert knowledge or previous studies, and an option is also to use random sampling without replacement, but in this latter case, there are risks regarding negative class labeling [Golini, 2011]. In the work of Marsh *et al*. [2023], the SDM profiling method was proposed, which can add pseudo-negative samples by analyzing the sensitivity of unlabeled instances in the model, examining the interaction of environmental conditions on species occurrence probability response curves.

From the conducted literature review, it can be concluded that the application of the Data-Centric Artificial Intelligence (DCAI) approach, with Confident Learning (CL) techniques for handling uncertainties and cleaning the dataset used for Species Distribution Modeling tasks, was unprecedented in the literature, with no other works identified with similar scope, until the work of Miyaji *et al*. [2023] was published. As an extension of that paper, this work proposes a new method - the Ensemble Confident Learning (ECL) - that is the first one to use multiple learners trained in different datasets resulting from Imbalanced Learning techniques to address the risk of biased results from the Meta-Learning process of Confident Learning.

3 Methods

In this section, the Confident Learning and Ensemble Confident Learning techniques are presented, along with the methodological procedure adopted to apply them in the case study.

3.1 Confident Learning

The Confident Learning (CL) technique proposed by Northcutt *et al*. [2021b] fall within the realm of Supervised Learning, capable of characterizing uncertain labels, identifying instances where they occur, using them for learning, and identifying ontology-related label issues.

The method is based on three principles: Noisy Data Pruning aims to search for and identify errors in labels; Count for uncertainty estimation, avoiding error propagation in model weights with imperfect probabilities; and Rank instances according to their uncertainty estimate, to select instances used for model training, enabling training with greater confidence [Northcutt *et al*., 2021b].

To achieve this, the method seeks to estimate the joint probability distribution of true labels and noisy labels in the analyzed dataset, estimating a matrix where rows represent noisy labels and columns represent true labels, filled with instance counts. Thus, the matrix's diagonals indicate instances where noisy and true labels are equal.

The joint probability distribution of true and noisy labels, represented by $Q_{\tilde{y},y^*}[i][j]$, is obtained from the matrix with instance counts in each class, represented by $C_{\tilde{y},y^*}[i][j]$, and its posterior normalization. In the $C_{\tilde{y},y^*}[i][j]$ matrix, $\tilde{y}$ represents noisy labels, $y*$ represents true labels, and i, j represent the rows and columns of the matrix, respectively.

The estimation of the joint probability distribution of true and noisy labels is presented in Equation (1). The definition of $\hat{X}_{\tilde{y}=i,y^*=j}$ is presented in Equation (2), where X represents the dataset, $\hat{p}$ is the estimated or predicted probability for the label, x is the instance, θ are the parameters of the Machine Learning model, t_j is the expected self-confidence for class j, and M is the set of all possible classes.

$$C_{\tilde{y},y^*}[i][j] := |\hat{X}_{\tilde{y}=i,y^*=j}| \tag{1}$$

$$\hat{X}_{\tilde{y}=i,y^*=j} := \{x \in \hat{X}_{\tilde{y}=i} : \hat{p}(\tilde{y}=j;x,\theta) \geq t_j, \\ j = argmax_{k\in M:\hat{p}(\tilde{y}=j;x,\theta)\geq t_j}\hat{p}(\tilde{y}=j;x,\theta)\} \tag{2}$$

Through Equation (2), it can be observed that to determine the joint probability distribution a comparison between the estimated probability of an instance belonging to a class labeled j with the expected self-confidence for class t_j is done. The self-confidence is calculated by Equation (3). Thus, an instance is considered for counting (and therefore can be confidently assigned to that class) only when its estimated probability is greater than or equal to the classification threshold established for the class (expected self-confidence for class t_j). This concept proposed by Northcutt *et al.* [2021b] is capable of generalizing the proposition of [Elcan and Noto, 2008].

$$t_j = \frac{1}{|\hat{X}_{\tilde{y}=j}|}\sum_{x\in\hat{X}_{\tilde{y}=j}}\hat{p}(\tilde{y}=j;x,\theta) \tag{3}$$

After determining the matrix $C_{\tilde{y},y^*}[i][j]$, it should be normalized to obtain the estimate of the distribution $Q_{\tilde{y},y^*}[i][j]$. Through it, it is possible to identify the off-diagonal markings, which indicate labels with higher associated uncertainty and are possibly incorrect (i.e., where the noisy label $\tilde{y}$ is different from the true label y^* with an estimated probability above the threshold t_j).

As advantages of using the Confident Learning (CL) technique, the following can be mentioned: the method does not have hyper parameters; it utilizes the Cross-Validation process to obtain probabilities; it can estimate the joint probability distribution directly from true and noisy labels; it can be used for multi-class classification problems; it can identify and rank instances according to their uncertainty and probability of being incorrect; it does not assume an uniform error distribution among classes; it is model-agnostic, applicable to any Machine Learning Classifier; and it does not require only instances with completely correct labels [Northcutt *et al.*, 2021b].

The steps of the presented method can be observed in Figure 1. Initially, noisy data feed the predictive model, which predicts their true labels y^* and calculates the estimated probabilities. Then, these are compared with the noisy labels $\tilde{y}$, obtaining the matrix $C_{\tilde{y},y^*}[i][j]$ and the joint probability distribution $Q_{\tilde{y},y^*}[i][j]$. This allows for pruning, identifying data with incorrect labels and generating a treated dataset.

3.2 Ensemble Confident Learning

The proposed method - Ensemble Confident Learning (ECL) - is based on Confident Learning [Northcutt *et al.*, 2021b]. The main difference between the methods is in the selection of instances with error in its label. While in the original method, the estimated probability of an instance belonging to each class j is compared with the trained model's expected self-confidence for each class t_j and the same model is afterwards used for prediction, in ECL the concept of Ensemble

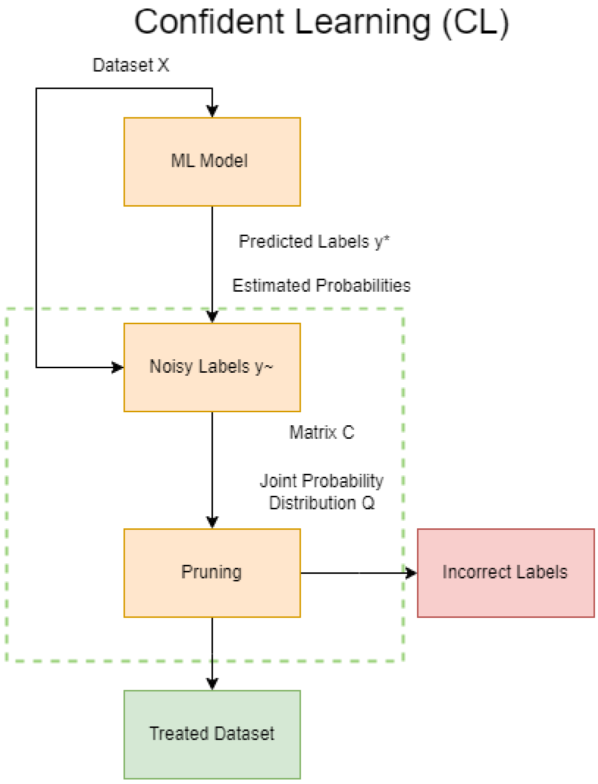

Figure 1. *Confident Learning* (CL). Adapted from Northcutt *et al.* [2021b]

Learning [James *et al.*, 2013] is used to reduce the risk of bias.

The proposed algorithm for ECL is presented bellow (Algorithm 1). The steps of the presented method can be observed in Figure 2. In ECL, multiple Machine Learning models are trained to determine the estimated probability of an instance belonging to each class j and the expected self-confidence for each class t_j. Inspired by the processes of Balanced Mini-Batch Training of Neural Networks [Shimizu *et al.*, 2018] and Co-teaching [Han *et al.*, 2018], multiple datasets with different frequencies of samples belonging to the minority and majority classes are generated from the original dataset with Imbalanced Learning techniques.

Initially, the amount of Machine Learning Classifiers n used for ECL and the different ratios α_i between minority and majority classes are defined. The ratios can be defined through a simple rule, such as 1:1 to 1:n, when n Classifiers are used. In case of multi-class Classification, a similar rule can be used, such as 1:1:1, 1:2:1, 1:1:2 for 3 classes and 3 Classifiers.

Then, for each Machine Learning Classifier C_i, the Imbalanced Learning technique is applied to the Noisy Dataset X to generate a new one with ratio of α_i (X_i), as presented in lines 1 to 4 of Algorithm 1. One possible approach that can be used is oversampling the minority class with Synthetic Minority Oversampling Technique (SMOTE) or other methods. SMOTE is a method of balancing the dataset through resampling. It involves creating synthetic positive samples, increasing their frequency. Another possible approach is undersampling instances of the majority class with techniques, such as Random Undersampling.

The Machine Learning Classifier C_i fits the dataset X_i. Then, as it happens in Confident Learning, for each instance x_l of the dataset X_i and each class j, the Classifier C_i is used to generate cross-validated estimates $\hat{p}(\tilde{y} = j; x_l, \theta)$, as presented in lines 5 to 10 of Algorithm 1. These estimated probabilities $\hat{p}(\tilde{y} = j; x_l, \theta)$ are used to determine the expected self-confidence for each class t_j, as presented in lines 11 to 14 of Algorithm 1. Then, the estimate $\hat{p}(\tilde{y} = j; x_l, \theta)$ for each instance and each class is compared to the self-confidence for each class t_j to determine the correct label y^*, as presented in lines 15 to 23 of Algorithm 1. With y^* and $\tilde{y}$, it is possible to determine matrices $C_{\tilde{y},y^*}[i][j]$ with instance counts in each class and joint probability distributions of true and noisy labels $Q_{\tilde{y},y^*}[i][j]$.

Finally, differently from the original method, in ECL to prune the data and mark an instance as correct, a voting between the three Classifiers is performed. An unanimous vote between the three Classifiers must be reached to classify an instance label as correct, i.e. $\hat{p}(\tilde{y} = j; x, \theta) \geq t_j$ for each model, as presented in lines 24 to 31 of Algorithm 1.

ECL relies on the assumption that datasets with different class predominances enable Machine Learning Classifiers to more accurately capture the relationships between the predictor and the target variables for each of the classes in the dataset [The Imbalanced-learn Developers, 2024]. Furthermore, the unanimous voting process between the different Classifiers minimizes the risk of biased results when selecting the instances with incorrect labels [James *et al.*, 2013].

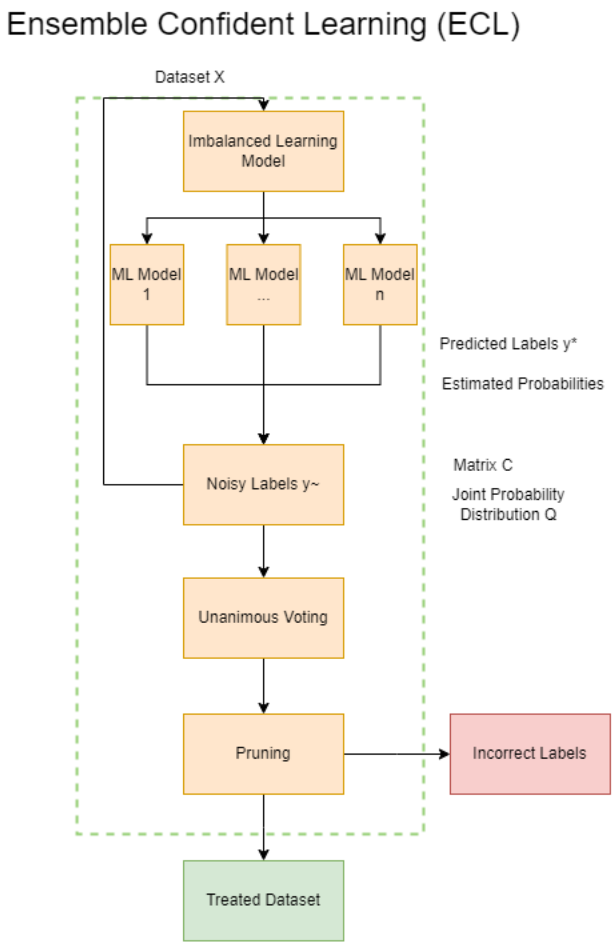

Figure 2. *Ensemble Confident Learning (ECL)*

Algorithm 1: Ensemble Confident Learning

Input : n Machine Learning Classifiers C, J classes, Imbalanced Learning Method L and ratio between minority and majority classes α
Data : Noisy dataset X with size N

1 # Part 1: Generate different datasets and train Classifier;
2 **for** $i = 1$ *to* n **do**
3 $X_i \leftarrow$ fit imbalanced learning($L(X, \alpha_i)$);
4 fit classifier($C_i(X_i, \tilde{y})$);
5 # Part 2: Estimate probabilities with Cross Validation;
6 **for** $l = 1$ *to* N **do**
7 **for** $j = 1$ *to* J **do**
8 $\hat{p}(\tilde{y} = j; x_l, \theta) \leftarrow C_i$ predict crossval prob;
9 **end**
10 **end**
11 # Part 3: Calculate threshold for each class;
12 **for** $j = 1$ *to* J **do**
13 $t_j \leftarrow$ calculate threshold($\hat{p}(\tilde{y} = j; x, \theta)$);
14 **end**
15 # Part 4: Predict label y^*;
16 **for** $l = 1$ *to* N **do**
17 **for** $j = 1$ *to* J **do**
18 **if** $\hat{p}(\tilde{y} = j; x, \theta) \geq t_j$ **then**
19 $y^* \leftarrow j$;
20 **end**
21 **end**
22 **end**
23 **end**
24 # Part 5: Prune data with Unanimous Voting;
25 **for** $l = 1$ *to* N **do**
26 **for** $i = 1$ *to* n **do**
27 **if** $y^* \neq \tilde{y}$ **then**
28 mark as incorrect(x_l,$\tilde{y}_l$);
29 **end**
30 **end**
31 **end**

3.3 Case Study

The Case Study presented in this paper is regarding a Species Distribution Modeling experiment using Machine Learning Classifiers, as it involves a problem with uncertain data [Martin *et al*., 2005].

For this purpose, the selected region for analysis was the Amazon Rain forest, more specifically, the city of Manaus (AM). This area is considered by experts as an ideal laboratory for studying the influence of anthropogenic activity on climate and terrestrial ecosystems in tropical forests [Martin *et al*., 2017]. Thus, between 2014 and 2015, meteorological and aerosol data were collected through low altitude flights. These flights were part of the Green Ocean Amazon 2014/15 (GoAmazon 2014/15) project, organized by the Atmospheric Radiation Measurement (ARM), an agency affiliated with the United States government, and Brazilian institutions [Martin *et al*., 2016]. From the GoAmazon 2014/15 project dataset, it was possible to obtain the predictor variables to be used for Species Distribution Modeling.

As for the data used as the target variable related to species occurrence, it was collected from two sources: the Chico Mendes Institute for Biodiversity Conservation (ICMBio), which monitors national biodiversity and openly provides species occurrence records throughout Brazil on the Biodiversity Portal [ICMBio, 2024]. In addition to the Global Biodiversity Information Facility (GBIF), one of the largest repositories of data on species occurrence on all continents [GBIF, 2024]. However, for the region analyzed in the case study, when combining both datasets, a low quantity of data for the same species is obtained. This fact may be prejudicial to the performance of Machine Learning models [Almeida *et al*., 2021].

To construct the dataset for Species Distribution Modeling, the data resulting from the spatial interpolation process, applied to the GoAmazon 2014/15 project data and made available by Miyaji *et al*. [2021], was used. The variables provided were: temperature, ozone concentration (O_3), carbon monoxide concentration (CO), nitrogen oxides concentration (NO_X), methane concentration (CH_4), carbon dioxide concentration (CO_2), isoprene concentration, acetonitrile concentration, particle number count, and water vapor volume fraction (H_2O).

For the same period of interest and the region between the same geographical coordinates of the spatial interpolation data, occurrence records of species provided by ICMBio and GBIF were obtained. Thus, the species with the highest occurrence counts were the *Coragyps atratus* (black vulture) and the *Tyrannus melancholicus* (tropical kingbird), representing 54 and 50 distinct records, respectively.

Using Python language, the bioclimatic dataset was constructed, applying the necessary filters, treatments, and performing the join operation between the datasets, considering the geographical coordinates and occurrence registration date. The species to be analyzed was the one with the highest occurrence frequency, the *Coragyps atratus*, to facilitate the application of Species Distribution Models [Hernandez *et al*., 2006].

Then, to select the predictor variables to be used in the Species Distribution Model, a Correlation Analysis was performed. Adopting the Pearson coefficient, the linear relationship between variables in pairs was analyzed. We decided to remove one of the variables from pairs with high correlation, i.e., with a Pearson coefficient greater than or equal to 80% in modulus [Mateo *et al*., 2013]. Thus, it was avoided that the model incorporated random patterns and suffered from multicollinearity [Pinaya and Corrêa, 2014]. Three predictor variables were removed from the dataset. Thus, the predictor variables used were: maximum temperature, minimum temperature, ozone concentration, carbon monoxide concentration, nitrogen oxides concentration, methane concentration, isoprene concentration, acetonitrile concentration, and water vapor volume fraction.

The selected Classification model was one of the most common for the task of Species Distribution Modeling: Logistic Regression, as it is a simple linear model and one of the most frequent in the literature with greater potential [Hegel *et al*., 2010; Beery *et al*., 2021].

Due to the natural class imbalance and the uncertainties associated with the labeling of negative absence classes of the species occurrence [Johnson *et al*., 2012], to treat the problem of Imbalanced Classification, the Synthetic Minority Oversampling Technique (SMOTE) [The Imbalanced-learn Developers, 2024] was chosen to be applied for both techniques: Confident Learning (CL) and Ensemble Confident Learning (ECL). For the case study of the original model (CL), it was defined that the resampled dataset should have a ratio of 1:3 between positive over negative samples. The size of the dataset was of 185,305 rows. In the original dataset, the sample distribution of different classes was 3.3% for the positive class (species presence) and 96.7% for the negative class (species absence). In the resampled dataset using the SMOTE technique, the distribution was 23% for the positive class and 77% for the negative class.

For the case study using Ensemble Confident Learning, the resampled datasets were generated from the original one through SMOTE. The ratios between the minority and majority samples were defined according to the rule 1:1 to 1:n, when n Classifiers are used.

With the dataset and the Classifiers defined, it was possible to apply Confident Learning (CL) [Northcutt *et al*., 2021b] and Ensemble Confident Learning (ECL) techniques, according to the procedure presented in Figures 1 and 2.

The other state-of-the-art Confident Learning based methods proposed by Zhang *et al*. [2023] and Li *et al*. [2023] were also applied to the case study, in order to provide different baselines. The Decoupled Confident Learning (DeCoLe) [Li *et al*., 2023] method was applied with two Classifiers trained in different datasets. The original training dataset was divided into two, according to a variable that is capable of characterizing different groups. For the case study, the selected variable was the geographic region: close to cities, such as Manaus (AM), and far from them. The Fairness Confident Learning [Zhang *et al*., 2023] method was applied to the case study with two Logistic Regression models with different hyper parameters and therefore learning capabilities.

As evaluation metrics for the Classifiers, Accuracy and Recall were adopted, as it is a problem of Imbalanced Classification. Furthermore, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was also evaluated. The metrics were evaluated on an hold-out dataset, using Cross Validation with the Stratified K-Fold method, with $K = 5$, as recommended for the application of the Confident Learning technique [Northcutt *et al.*, 2021b].

The experimental protocol for the Case Study is presented in Figure 3.

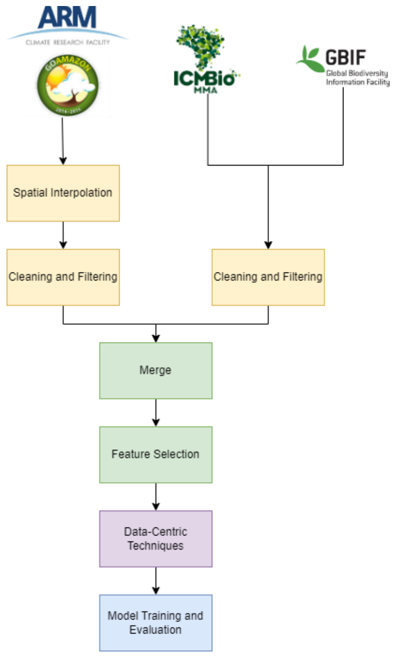

Figure 3. Experimental protocol for the Case Study

4 Results

The Confident Learning (CL) technique was applied using the Logistic Regression model for the case study. Initially, it was applied to the original dataset, modified only by incorporating synthetic positive samples through the SMOTE resampling technique. Through its use, the predictive capability of the classifier could be improved, especially regarding the minority class. However, due to the creation of synthetic samples, larger uncertainties may be introduced into the dataset. In this sense, using techniques such as Confident Learning (CL) becomes even more relevant.

For the optimization of the model's hyper parameters, Cross Validation was used, considering the parameter C. This corresponds to the inverse of regularization, which controls the model's robustness to small data variations, preventing overfitting [James *et al.*, 2013]. After this process, the ideal value was determined to be $C = 1$.

Then, Cross Validation was applied using the 5-Fold method to measure the defined evaluation metrics, obtaining an average accuracy of 76.7%, a recall of 50.0% for the positive (minority) class, and a ROC-AUC of 78.1%. Thus, it is noted that the model was able to develop good predictive capability, but this does not reflect when considering only the minority class, which has greater importance for the Species Distribution Modeling task.

The trained model was used for the Confident Learning technique. For this purpose, the classification threshold values were determined, i.e., the expected self-confidence for each of the classes t_j. The values of $t_0 = 79.9\%$ for the negative class and $t_1 = 33.1\%$ for the positive class were obtained. This indicates that the model has lower self-confidence for the minority class.

With the expected self-confidence defined, it was possible to construct the joint matrix $C_{\tilde{y},y^*}[i][j]$. Then, the values were normalized, obtaining the joint probability distribution $Q_{\tilde{y},y^*}[i][j]$, presented in Table 2. Approximately 24.0% of the dataset has incorrect labels according to the CL method.

Table 2. Joint Probability Distribution $Q_{\tilde{y},y^*}[i][j]$ for CL

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	62.4%	16.7%
$\tilde{y} = 1$	7.3%	13.6%

For the Ensemble Confident Learning technique, n different Logistic Regression Classifiers were trained each in one of the resampled datasets with ratios between the minority and majority samples of 1:1 to 1:n, using the same procedure described above. The method was applied with $n = 3$ Classifiers (1:1, 1:2 and 1:3 ratios), $n = 4$ Classifiers (1:1, 1:2, 1:3 and 1:4 ratios) and $n = 5$ Classifiers (1:1, 1:2, 1:3, 1:4 and 1:5 ratios). The intermediate results of the ECL method will be presented for the method with $n = 3$ Classifiers. The expected self-confidence t_j for each of the classes are presented in Table 3. For the Classifier trained in the dataset with ratio of 1:3, the self-confidence for the majority class is the highest, while for the minority class is the lowest. Using only this Classifier could be harmful for the CL performance, since it has a low predictive ability for the minority class, which could lead to a large amount of instances being mislabeled and, therefore, increasing the risk of obtaining biased results [Zhang *et al.*, 2023]. As the ratio between minority and majority classes increases, the self-confidence in the minority class of the models trained in the resampled dataset increases up to 62.0% for 1:1 ratio. However, meanwhile the self-confidence in the majority class decreases, but only 17.9 p.p.. For the model trained in the resampled dataset with 1:1 ratio, the self-confidence for both classes are high (62.0%).

Table 3. Expected self-confidence for each class for each Classifier for ECL with $n = 3$ Classifiers

Ratio	t_0	t_1
1:3	79.9%	33.1%
1:2	66.3%	55.1%
1:1	62.0%	62.0%

Then, the joint probability distribution $Q_{\tilde{y},y^*}[i][j]$ was computed for each of the three models, presented in Tables 4, 5 and 6. Approximately 24.0% of the dataset has incorrect labels according to the model with ratio 1:3, 18.3% for the model with ratio 1:2 and 17.5% with ratio 1:1. As the ratio increases, the error rate decreases. However, a caveat should be highlighted: since SMOTE is used to resample the datasets, the total amount of instances are different for each of the models.

Then, the dataset was pruned according to the classifications presented in Table 2 for the CL method. Then, a new

Table 4. Joint Probability Distribution $Q_{\tilde{y},y^*}[i][j]$ for ECL with Ratio 1:3 for ECL with $n = 3$ Classifiers

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	62.4%	16.7%
$\tilde{y} = 1$	7.3%	13.6%

Table 5. Joint Probability Distribution $Q_{\tilde{y},y^*}[i][j]$ for ECL with Ratio 1:2 for ECL with $n = 3$ Classifiers

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	50.5%	16.2%
$\tilde{y} = 1$	2.1%	31.1%

Table 6. Joint Probability Distribution $Q_{\tilde{y},y^*}[i][j]$ for ECL with Ratio 1:1 for ECL with $n = 3$ Classifiers

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	44.8%	15.3%
$\tilde{y} = 1$	2.2%	37.7%

Logistic Regression model was trained in the treated dataset. With the new model, Cross Validation was again applied using the 5-Fold method to obtain evaluation metrics. An average accuracy of 95.6%, a recall of 75.0% for the positive (minority) class, and a ROC-AUC of 95.8% were obtained. Therefore, a significant improvement in the predictive capability of the model is observed in all metrics. Specifically, this enhancement was higher for metrics of the positive (minority) class, with an increase of 25 p.p. or 50% in recall.

For the ECL method with $n = 3$ Classifiers, the data pruning was performed with Unanimous Voting. The joint probability distribution $Q_{\tilde{y},y^*}[i][j]$ is presented in Table 7. Thus, the error rate increases to 33.9%, since it is a more conservative method. The same procedure described above was performed to train the new model, obtaining an average accuracy of 93.8%, a recall of 80.0% for the positive (minority) class, and a ROC-AUC of 93.8%. The comparison between the performance metrics of the models is presented in Table 8.

Table 7. Joint Probability Distribution $Q_{\tilde{y},y^*}[i][j]$ for ECL with Unanimous Voting for ECL with $n = 3$ Classifiers

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	60.3%	18.1%
$\tilde{y} = 1$	15.8%	5.8%

Table 8. Comparison of performance between models trained in the original and treated datasets

Dataset	Accuracy	Recall	ROC-AUC
Original	76.7%	50.0%	78.1%
CL	95.6%	75.0%	95.8%
DeCoLe	93.0%	80.0%	82.8%
Fairness CL	95.0%	85.0%	90.5%
ECL 3 Classifiers	93.8%	80.0%	93.8%
ECL 4 Classifiers	93.8%	90.0%	96.4%
ECL 5 Classifiers	93.9%	90.0%	99.1%

The improved performance metrics achieved with both Confident Learning (CL) and Ensemble Confident Learning (ECL) methods are justified by the fact that the methods are capable of cleaning the dataset. This is achieved by identifying instances where there is greater uncertainty about their label. However, there is a significant dependency on the expected self-confidence for each class, which determines whether an instance is considered as incorrect. Therefore, it is relevant for the method that the base classifier has high predictive capability. If this condition is not met, i.e., if the predictive capability is very low, this criterion becomes less restrictive, resulting in many instances being identified as incorrect, which can introduce biases to the desired analysis.

Comparing the performance metrics from Ensemble Confident Learning to the original method, for ECL with 3 Machine Learning Classifiers, there is a slight decrease in both Accuracy and ROC-AUC of less than 2 p.p.. In the other hand, there is an increase of 5 p.p. in the metric of the minority class, since Recall reaches 80.0%.

The performance of the ECL method with 4 Classifiers was better, with a ROC-AUC of 96.4% (0.6 p.p. higher than CL) and a Recall of 90% (15 p.p. higher than CL). The Accuracy was still slightly lower: 93.8% (1.8 p.p. lower than CL). Finally, the ECL method with 5 Classifiers was able to achieve the best performance, with a ROC-AUC of 99.1% (3.3 p.p. higher than the original method) and a Recall of 90%.

As presented in Table 8, there is a clear trend between the model's performance and the amount of Machine Learning Classifiers used for the ECL method. With the use of more Classifiers, the percentage of samples identified with incorrect labels increases and the model becomes more conservative, as an Unanimous Voting is adopted. The key assumption is that Classifiers trained in different datasets are able to develop unique predictive capabilities, which enhances their performance and ability to identify instances with incorrect labels. Therefore, with 3 Classifiers a slight increase in Recall is observed. With 4 or 5 Classifiers, a significant increase in performance is achieved in both Recall and ROC-AUC.

The results presented in Table 8 show that both methods DeCoLe and Fairness Confident Learning were able to achieve a higher Recall than the original method. However, there is a decrease in both Accuracy and ROC-AUC. The performance of Fairness Confident Learning is similar to Ensemble Confident Learning with 3 Classifiers. Thus, the use of multiple Classifiers improves the model's predictive capacity related to the minority class (higher Recall). However, to achieve an overall better performance more than 3 Classifiers are necessary.

Considering that ECL has a more robust process to select instances with incorrect label, since it is based in different Classifiers with higher self-confidence for both classes, it can be considered as a robust technique to address the problem of Label Noise Classification in Machine Learning.

5 Conclusion

In this paper, techniques from the field of Data-Centric Artificial Intelligence (DCAI) were applied to clean data, treat associated uncertainties, and enhance the performance of Machine Learning Classifiers. Specifically, a new method was proposed and evaluated - Ensemble Confident Learning (ECL). Furthermore, the original method - Confident Learning (CL) - was applied to identify instances in the original dataset there where more likely to be incorrect.

For this purpose, CL, ECL and other state-of-the-art techniques were applied to a case study of a Species Distribution Modeling experiment. This task involves various sources of uncertainties associated with the labels of the target variable (the presence or absence of the analyzed species), as only species presence records (Presence-only Data) are used to construct the dataset. The dataset adopted regarded the occurrence of *Coragyps atratus*, with predictor variables related to meteorological and aerosol data.

Due to the class imbalance in the dataset used for modeling, the Synthetic Minority Oversampling Technique (SMOTE) was applied to resample the positive class for the CL method. Subsequently, a Logistic Regression model was trained, and its performance metrics were calculated. This model was then used to apply the Confident Learning technique to identify instances with potentially incorrect labels and to clean the data. As a result, we observed an increase from 76.7% to 95.6% in Accuracy and from 78.1% to 95.8% in ROC-AUC. The most significant improvement was in the metric associated with the minority class, with an increase of 50% (25 p.p.) in Recall.

To apply ECL with 3, 4 and 5 Machine Learning Classifiers, SMOTE was also used to generate three resampled datasets with different ratios between minority and majority classes. A different Logistic Regression Classifier was trained in each of the datasets and the dataset was pruned with Unanimous Voting. The best performing model used 5 Classifiers and obtained an average Accuracy of 93.9%, a Recall of 90.0% for the positive (minority) class, and a ROC-AUC of 99.1%, showing a significant improvement of 20% in Recall and 3.5% in ROC-AUC in comparison with Confident Learning.

Therefore, it is concluded that both Confident Learning (CL) and Ensemble Confident Learning (ECL) techniques show great potential for treating uncertainties in datasets for Machine Learning Classification tasks, resulting in a significant improvement in performance. ECL can be considered as more robust, since it does not depend on an unique model to select instances with incorrect label. Thus, it reduces risk of biased outputs resulting from the Meta-Learning process of Confident Learning using multiple learners (Ensemble).

Given the obtained results, for future works, we suggest applying both Confident Learning (CL) and Ensemble Confident Learning (ECL) techniques to other commonly used Machine Learning classifiers for Species Distribution Modeling, such as Random Forests, Support Vector Machines, Artificial Neural Networks, among others. In these cases, as the model's predictive ability may be higher, the criterion adopted to evaluate whether an instance has an incorrect label can become more accurate, leading to improved results. Different learners composing the ensemble could be applied in the ECL method, which could enhance the identification of instances with noisy labels. Moreover, another suggestion for future works could be the use of different types of voting. In ECL, a more conservative approach was adopted through Unanimous Voting.

Funding

We would like to specially thanks the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for funding the project through a scholarship from the PIBIC program (2020/21 - 1745), and the Fundação de Amparo à Pesquisa e Inovação do Estado de São Paulo (FAPESP) for funding the project through the Thematic Projects "Life cycles and aerosol clouds in the Amazon" (2017/17047-0) and "Research Centre for Greenhouse Gas Innovation - RCG2I" (2020/15230-5), and the researchers from the Big Data Research Group and Data Science at EPUSP.

Authors' Contributions

Conceptualization — R.O.M.; Methodology — R.O.M.; Experiments — R.O.M.; Writing — Original Draft Preparation — R.O.M. and F.V.d.A.; Writing — Review and Editing — P.L.P.C.; Visualization — R.O.M. All authors have read and agreed to the final version of the manuscript.

References

Almeida, F. V., Bueno, W. M., Miyaji, R. O., and Corrêa, P. L. P. (2021). Experimento de modelagem de distribuição de espécies baseada em variáveis ambientais e de aerossóis na região próxima a manaus (am). In *Anais do XII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*. SBC.

Beery, S., Cole, E., Parker, J., Perona, P., and Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review. In *Proceedings of ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) 2021*.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of 26th International Conference on Machine Learning*. ACM.

Di Lorenzo, B., Farcomeni, A., and Golini, N. (2011). A bayesian model for presence-only semicontinuous data, with application to prediction of abundance of taxus baccata in two italian regions. *Journal of Agriculture Biological and Environmental Statistics*, 16:339–356.

Elcan, K. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*.

Elcan, K. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2008*.

Forman, G. (2005). Counting positives accurately despite inaccurate classification. In *Proceedings of the 16th European Conference on Machine Learning*.

GBIF (2024). Gbif | global biodiversity information facility. https://www.gbif.org/. Online; accessed 09-March-2024.

Golini, N. (2011). *Bayesian Modelling of Presence-only Data*. PhD thesis, Spienza Universidade de Roma.

Hamid, O. H. (2022). From model-centric to data-centric ai: A paradigm shift or rather a complementary approach? In *Proceedings of 2022 8th International Conference on Information Technology Trends (ITT)*, pages 45–54. IEE.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceeding of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Hegel, T. M., Cushman, A., Evans, J., and Huetmann, F. (2010). *Spatial Complexity, Informatics and Wildlife Conservation*, chapter Current State of the Art for Statistical Modelling of Species Distributions. Springer.

Hernandez, P. A., Graham, C. H., Master, L. L., and Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5):773–785. DOI: https://doi.org/10.1111/j.0906-7590.2006.04700.x.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Huang, J., Qu, L., Jia, R., and Zhao, B. (2019). O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the International Conference on Computer Vision (ICCV) 2019*.

Hutchinson, G. E. (1991). Population studies: Animal ecology and demography. *Bulletin of Mathematical Biology*, 53(1-2):193–213.

ICMBio (2024). Portal da biodiversidade do instituto chico mendes de conservação da biodiversidade. https://portaldabiodiversidade.icmbio.gov.br/portal/. Online; accessed 09-March-2024.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, Londres.

Johnson, R., Chawla, N., and Hellmann, J. (2012). Species distribution modeling and prediction: A class imbalance problem. pages 9–16. DOI: 10.1109/CIDU.2012.6382186.

Li, Y., De-Arteaga, M., and Saar-Tsechansky, M. (2023). Mitigating label bias via decoupled confident learning. In *Proceeding of the AI HCI Workshop at the 40th International Conference on Machine Learning (ICML)*.

Lipton, Z., Wang, Y., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the International Conference on Machine Learning (ICML) 2018*.

Marsh, J. C., Gavish, Y., Kuemmerlen, M. C., Stoll, S., Haase, P., and Kunin, W. E. (2023). Sdm profiling: A tool for assessing the information-content of sampled and unsampled locations for species distribution models. *Ecological Modelling*, 475(1).

Martin, S. T., Artaxo, P., Machado, L., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Biscaro, T., Brito, J., Calheiros, A., *et al*. (2017). The green ocean amazon experiment (goamazon2014/5) observes pollution affecting gases, aerosols, clouds, and rainfall over the rain forest. *Bulletin of the American Meteorological Society*, 98(5):981–997.

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Andreae, M. O., Barbosa, H., Fan, J., *et al*. (2016). Introduction: observations and modeling of the green ocean amazon (goamazon2014/5). *Atmospheric Chemistry and Physics*, 16(8):4785–4797.

Martin, T. G., Kuhnert, P. M., Mengersen, K., and Possingham, H. P. (2005). The power of expert opinion in ecological models using bayesian methods: Impact of grazing on birds. *Ecological Applications*, 15:266–280.

Mateo, R. G., Vanderpoorten, A., Muñoz, J., Laenen, B., and Désamoré, A. (2013). Modeling species distributions from heterogeneous data for the biogeographic regionalization of the european bryophyte flora. *PLoS One*, 8(2):e55648.

Miyaji, R., Almeida, F., and Corrêa, P. (2023). Aplicação de técnicas de confident learning para limpeza de dados e melhoria de desempenho de classificadores de aprendizado de máquina: um estudo de caso. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*.

Miyaji, R. O., Bauer, L. O., Ferrari, V. M., Almeida, F. V., Corrêa, P. L. P., and Rizzo, L. V. (2021). Interpolação espacial de variáveis ambientais e aerossóis na região da bacia amazônica próxima a manaus-am. In *Anais do XII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*. SBC.

Miyaji, R. O. and Corrêa, P. L. P. (2021). Handling uncertainty through bayesian inference for species distribution modelling in the amazon basin region. In *2021: ANAIS DO XVIII ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL*.

Northcutt, C. G., Athalye, A., and Mueller, J. (2021a). Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Northcutt, C. G., Jiang, L., and Chuang, I. L. (2021b). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research (JAIR)*, 70(1):1373–1411.

Pinaya, J. and Corrêa, P. (2014). Metodologia para definição das atividades do processo de modelagem de distribuição de espécies. In *Anais do V Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais*, pages 45–54, Porto Alegre, RS, Brasil. SBC.

Shimizu, R., Asako, K., Ojima, H., Morinaga, S., Hamada, M., and Kuroda, T. (2018). Balanced mini-batch training for imbalanced image data classification with neural network. In *Proceeding of the First International Conference on Artificial Intelligence for Industries (AI4I)*.

The Imbalanced-learn Developers (2024). Imbalanced-learn documentation. `https://imbalanced-learn.org/stable/`. Online; accessed 09-March-2024.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Zhang, Y., Li, B., Ling, Z., and Zhou, F. (2023). Mitigating label bias in machine learning: Fairness through confident learning. *arXiv*, 2312.08749. DOI: https://doi.org/10.48550/arXiv.2312.08749.