Aplicação de Técnicas de *Confident Learning* para Limpeza de Dados e Melhoria de Desempenho de Classificadores de Aprendizado de Máquina: um Estudo de Caso

Renato O. Miyaji[1], **Felipe V. de Almeida**[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{re.miyaji,felipe.valencia.almeida,pedro.correa}@usp.br

Abstract. *Model-Centric techniques, such as hyperparameter selection and regularization, are commonly used in the literature to improve the performance of Machine Learning Classifiers. However, when a dataset with uncertain data is used, Data-Centric approaches have a good potential. These methods aim to systematically engineer data to improve model performance. Thus, Confident Learning (CL) techniques were applied for a study case of Species Distribution Modeling in the Amazon Basin using Machine Learning Classifiers, which aimed to predict the probability of occurrence of a species, given environmental conditions. In comparison with Model-Centric methods, CL techniques presented a 23% improvement of ROC-AUC for Logistic Regression.*

Resumo. *Técnicas centradas em modelos, como otimização de hiper parâmetros e regularizações, são comumente utilizadas na literatura para aprimorar o desempenho de Classificadores de Aprendizado de Máquina. Entretanto, quando tratando um conjunto de dados com incertezas, abordagens centradas em dados apresentam bom potencial. Assim, técnicas de Confident Learning (CL) foram aplicadas para um estudo de caso de Modelagem de Distribuição de Espécies na Amazônica utilizando Classificadores para estimar a probabilidade de ocorrência de uma espécie, com base em condições ambientais. Em comparação com métodos centrados em modelos, as técnicas CL apresentaram uma melhoria de 23% no ROC-AUC para Regressão Logística.*

1. Introdução

Para aprimorar o desempenho de modelos de Aprendizado de Máquina, geralmente são aplicadas técnicas direcionadas a melhorar o processo de treinamento dos modelos ou a otimização de seus hiper parâmetros [James et al. 2013]. Exemplos desses métodos são as regularizações *Lasso* [Tibshirani 1996] e *Ridge* [Hoerl and Kennard 1970] no caso da técnica de Regressão Linear.

Na literatura, outras técnicas também foram desenvolvidas para essa finalidade. Porém, elas buscam modificar o conjunto de dados de treinamento de maneira sistemática, a fim de se aprimorar o desempenho dos modelos de Aprendizado de Máquina. Essa abordagem é nomeada de Inteligência Artificial centrada em Dados (*Data-Centric Artificial Intelligence* - DCAI) e se difere dos métodos mais comuns direcionados a melhoria dos modelos: a Inteligência Artificial centrada em Modelos (*Model-Centric Artificial Intelligence* - MCAI) [Hamid 2022].

A abordagem centrada em Dados (DCAI) mostra-se útil em diversos contextos, especialmente quando o conjunto de dados sobre o qual se realiza a modelagem é limitado - ou seja, coletar mais dados é uma tarefa complexa ou impossível - ou até quando existem incertezas em relação aos dados coletados [Hamid 2022]. Dessa forma, existem duas classes de técnicas: as que utilizam algoritmos para obter uma melhor compreensão sobre os dados e fazem o uso dessa informação para aprimorar o processo de treinamento dos modelos, como no caso de *Curriculum Learning* [Bengio et al. 2009]; e os algoritmos que modificam os dados de treinamento para melhorar o desempenho dos modelos de Aprendizado de Máquina [Northcutt et al. 2021b].

Uma dessas técnicas é a proposta por [Northcutt et al. 2021b], a *Confident Learning* (CL). Como um método DCAI, ela possui o foco em avaliar a qualidade dos rótulos da variável resposta (*Labels*) dos dados, através da caracterização e identificação de erros. Isso é realizado com base nos princípios de poda (*Pruning*) de dados incertos, com a definição de limites probabilísticos para estimar as incertezas e, então, ordenar as observações incertas e realizar o treinamento em um conjunto de dados com maior confiança.

As técnicas de *Confident Learning* foram utilizadas com sucesso na literatura [Northcutt et al. 2021a], sendo aplicadas sobre os 10 conjuntos de dados mais comuns na Ciência de Dados, usados para casos de visão computacional, processamento de linguagem natural, áudios, entre outros. Com o uso de CL, foi possível identificar uma média de 3,3% do total de observações com dados incertos que, quando tratados ou retirados, resultaram em um melhor desempenho dos modelos de Aprendizado de Máquina utilizados.

Nesse contexto, este trabalho buscou aplicar a abordagem de Inteligência Artificial centrada em Dados (*Data-Centric Artificial Intelligence* - DCAI), por meio das técnicas de *Confident Learning* (CL), para promover a limpeza do conjunto de dados e tratamento de incertezas contidades nele, a fim de se aprimorar o desempenho de Classificadores de Aprendizado de Máquina. Para isso, foi utilizado um estudo de caso a respeito de um experimento de Modelagem de Distribuição de Espécies (*Species Distribution Models* – SDM), no qual os Classificadores são utilizados para estimar a probabilidade de ocorrência de uma espécie com base nas variáveis ambientais.

Os modelos de Distribuição de Espécies são amplamente utilizados na Ecologia com o objetivo de se avaliar quantitativamente o nicho ecológico para a espécie analisada, ou seja, a faixa de valores das variáveis ambientais que tornam um habitat adequado para a ocorrência da espécie em estudo [Hutchinson 1991]. Nas últimas décadas, houve um grande avanço na área de Aprendizado de Máquina, com o desenvolvimento de modelos com desempenhos expressivos. Assim, esses passaram a ser utilizados com mais frequência para problemas de Modelagem de Distribuição de Espécies [Hegel et al. 2010].

Entretanto, os modelos de Aprendizado de Máquina apresentam um maior potencial para aplicações nas quais uma grande base de dados confiáveis está disponível para modelagem. Especificamente para a Modelagem de Distribuição de Espécies, este não é o caso, uma vez que são utilizados dados de ocorrência de espécies. Geralmente, esses são disponibilizados na forma de dados de apenas presença das espécies (*Presence-only Data*), ou seja, nos conjuntos de dados mais comuns constam apenas as observações de ocorrência da espécie - equivalente à classe positiva para o problema de Classificação -

ao passo que a afirmação de não ocorrência da espécie - equivalente à classe negativa - é um processo que incorpora incertezas aos dados. Uma prática comum de se aplicar é a incorporação de amostras pseudo-negativas (*Pseudonegatives* ou *Pseudoabsences*) para representar a classe negativa do conjunto de dados [Beery et al. 2021].

Porém, a utilização de dados não confiáveis para o processo de SDM, pode levar a resultados incorretos [Martin et al. 2005]. Nesse cenário, por permitir o tratamento de incertezas através da limpeza do conjunto de dados, as técnicas de *Confident Learning* (CL) mostram-se com grande potencial de aprimorar o uso de Classificadores de Aprendizado de Máquina para a Modelagem de Distribuição de Espécies.

2. Trabalhos Relacionados

A padronização de dados inconsistentes é um dos principais desafios no contexto de bancos de dados e aplicações relacionadas, como o desenvolvimento de análises e modelos preditivos. Por isso, diferentes técnicas para realizar a limpeza automática de conjuntos de dados e o tratamento de suas incertezas, a fim de viabilizar a aplicação de modelos de Classificação de Aprendizado de Máquina, foram desenvolvidas na literatura. Nos trabalhos de [Elcan 2001] e [Forman 2005], foram propostas técnicas que buscam estimar as taxas de falsos positivos e falsos negativos para a tarefa de Classificação Binária, como o método *Cost-Sensitive Learning*.

Já no trabalho de [Elcan and Noto 2008], foi proposta uma nova técnica para Classificação Binária mais robusta, uma vez que essa poderia ser aplicada para problemas nos quais se possui apenas dados a respeito da classe positiva, além de observações não rotuladas para a variável resposta. Isso foi feito, por meio da introdução do conceito de limite de classificação para permitir a identificação das observações incertas no conjunto de dados. No entanto, sua principal limitação se dava pela necessidade em se possuir dados da classe positiva totalmente confiáveis.

[Lipton et al. 2018] propõe a técnica *Black Box Shift Estimation* (BBSE) para identificar observações com rótulos invertidos ou incertos, através da utilização de Matrizes de Confusão e processos de Validação Cruzada. No trabalho de [Huang et al. 2019], é demonstrada a eficácia em se identificar as observações incertas, tratá-las e, então, realizar o treinamento do modelo de Aprendizado de Máquina sobre o conjunto de dados tratado, com um ganho significativo de desempenho.

A técnica de *Confident Learning* (CL), proposta por [Northcutt et al. 2021b], buscou incorporar as principais contribuições dos autores citados anteriormente, com o objetivo de se obter um método mais robusto e generalizável, capaz de lidar com observações incertas e/ou não rotuladas, identificá-las e tratá-las, sendo aplicável para qualquer tipo de modelo de Classificação de Aprendizado de Máquina, além de problemas de classificação com múltiplas classes.

Especificamente para a tarefa de Modelagem de Distribuição de Espécies, diversas abordagens foram adotadas na literatura para o tratamento das incertezas. Uma delas é através da abordagem Bayesiana, na qual é possível incorporar conhecimentos de especialistas ou estudos prévios aos modelos, por meio do Teorema de Bayes. Nos modelos Bayesianos, pode-se fornecer as distribuições de probabilidade *a priori* e de verossimilhança (*likelihood*) sobre seus parâmetros. A partir disso, determina-se a distribuição de

probabilidade *a posteriori*. O classificador de Regressão Logística Bayesiano, avaliado por [Miyaji and Corrêa 2021], [Di Lorenzo et al. 2011] e [Golini 2011], motrou-se com grande potencial para o tratamento das incertezas. Porém, a principal limitação dessa abordagem é a necessidade de conhecimento de especialistas ou estudos prévios para a sua aplicação, ademais ela não é aplicável a todos os modelos de classificação.

Outra possibilidade de abordagem é através da representação de classes negativas no conjunto de dados por meio de amostras pseudo-negativas (*Pseudonegatives*) [Beery et al. 2021]. Preferivelmente essas devem ser selecionadas a partir de conhecimentos de especialistas ou estudos prévios, sendo também uma opção a utilização de uma amostragem aleatória sem reposição, porém neste último caso existem riscos acerca da marcação da classe negativa [Golini 2011]. No trabalho de [Marsh et al. 2023], foi proposto o método *SDM profiling* que é capaz de adicionar amostras pseudo-negativas através de uma análise de sensibilidade das observações não rotuladas no modelo, analisando a interação das condições ambientais nas curvas de resposta de probabilidade de ocorrência da espécie.

A partir da revisão bibliográfica realizada, pôde-se concluir que a aplicação da abordagem de Inteligência Artificial centrada em Dados (DCAI), com as técnicas de *Confident Learning* (CL) para o tratamento das incertezas e limpeza do conjunto de dados utilizados para a tarefa de Modelagem de Distribuição de Espécies possui caráter inédito na literatura, não sendo identificados outros trabalhos com escopo semelhante.

3. Metodologia

A seguir, é apresentada a técnica *Confident Learning*, assim como o procedimento metodológico adotado para sua aplicação no estudo de caso.

3.1. *Confident Learning*

As técnicas de *Confident Learning* (CL) propostas por [Northcutt et al. 2021b] podem ser classificadas dentro da área de Aprendizado Supervisionado, sendo capazes de caracterizar rótulos incertos, identificar observações nas quais eles ocorram, utilizá-las para aprender e identificar problemas relacionados à ontologia dos rótulos.

O método possui como base três princípios: poda de dados incertos (*Noisy Data Pruning*) com o objetivo de se buscar e identificar erros nos rótulos; contagem (*Count*) para estimativa de incertezas e evitando propagação de erros nos pesos aprendidos pelo modelo com probabilidades imperfeitas; e ordenamento (*Rank*) das observações de acordo com sua estimativa de incerteza, de modo a selecionar as observações que serão utilizadas para o processo de treinamento do modelo, sendo possível treinar o modelo com maior confiança [Northcutt et al. 2021b].

Para realizar isso, o método busca estimar a distribuição de probabilidade conjunta dos rótulos verdadeiros (*True Labels*) e dos rótulos incertos (*Noisy Labels*) no conjunto de dados analisado, estimando uma matriz na qual nas linhas são representados os rótulos incertos e nas colunas os rótulos verdadeiros, sendo preenchida com a contagem das observações. Assim, as diagonais da matriz indicam as observações cujo rótulo incerto e rótulo verdadeiro são iguais.

A distribuição de probabilidade conjunta dos rótulos verdadeiros e incertos, representada por $Q_{\tilde{y},y^*}[i][j]$, é obtida a partir da matriz com a contagem de observações em

cada uma das classes, representada por $C_{\tilde{y},y^*}[i][j]$, e sua posterior normalização. Na matriz $C_{\tilde{y},y^*}[i][j]$, $\tilde{y}$ representa os rótulos incertos (*Noisy Labels*), $y*$ representa os rótulos verdadeiros (*True Labels*) e i, j representam as linhas e colunas da matriz, respectivamente.

A estimativa da distribuição de probabilidade conjunta dos rótulos verdadeiros e incertos é apresentada na equação (1). Já a definição de $\hat{X}_{\tilde{y}=i,y^*=j}$ é apresentada na equação (2), na qual X representa o conjunto de dados, $\hat{p}$ é a probabilidade estimada ou prevista para o rótulo, x é a observação, θ são os parâmetros do modelo de Aprendizado de Máquina, t_j é a auto-confiança esperada para a classe j e M é o conjunto de todas as classes possíveis.

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

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

Por meio da equação (2), nota-se que a determinação da distribuição de probabilidade conjunta é feita a partir da comparação entre a probabilidade estimada de uma certa observação pertencer a uma determinada classe com rótulo j e a auto-confiança esperada para a classe t_j, sendo essa calculada a partir da equação (3). Assim, uma observação é considerada para a contagem (e, portanto pode ser atribuída àquela classe de maneira confiável) apenas quando sua probabilidade estimada for maior ou igual ao limite de classificação estabelecido para a classe (auto-confiança esperada para a classe t_j), Esse conceito proposto por [Northcutt et al. 2021b] é capaz de generalizar a proposição de [Elcan and Noto 2008].

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

Após determinar a matriz $C_{\tilde{y},y^*}[i][j]$, deve-se normalizá-la para obter a estimativa da distribuição $Q_{\tilde{y},y^*}[i][j]$. A partir dela, pode-se identificar as marcações fora da diagonal, que indicam os rótulos com maior incerteza associada e que possivelmente são os incorretos (ou seja, cujo rótulo incerto $\tilde{y}$ é diferente do rótulo verdadeiro y^* com probabilidade estimada acima do limite t_j).

Como vantagem da utilização da técnica *Confident Learning* (CL), pode-se citar: o método não possui hiper parâmetros; utiliza o processo de Validação Cruzada para obter as probabilidades; é capaz de estimar a distribuição de probabilidade conjunta diretamente dos rótulos verdadeiros e incertos; pode ser utilizado para problemas de classificação com múltiplas classes; é capaz de identificar e ordenar as observações de acordo com seu nível de incerteza e probabilidade de estar incorreta; não assume uma distribuição uniforme de erro entre as classes; é agnóstica a modelo, sendo aplicável a qualquer modelo de Classificação; e não requer que existam apenas observações com rótulos totalmente corretos [Northcutt et al. 2021b].

As etapas do método apresentado podem ser observadas na Figura 1. Inicialmente, os dados incertos alimentam o modelo preditivo, que realiza a previsão de seus rótulos verdadeiros y^* e calcula as probabilidades estimadas. Em seguida, essas são confrontadas com os rótulos incertos $\tilde{y}$, obtendo a matriz $C_{\tilde{y},y^*}[i][j]$ e a distribuição de probabilidade conjunta $Q_{\tilde{y},y^*}[i][j]$. Isso permite a realização da poda, identificando os dados com rótulos incorretos e gerando um conjunto de dados tratado.

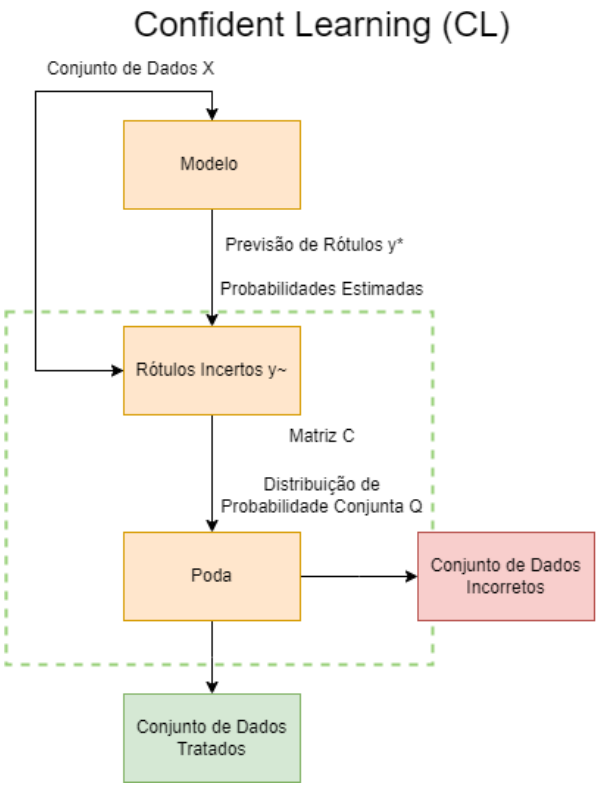

Figura 1. Método *Confident Learning* (CL). Adaptado de [Northcutt et al. 2021b]

3.2. Estudo de Caso

O Estudo de Caso selecionado foi em relação a um experimento de Modelagem de Distribuição de Espécies utilizando Classificadores de Aprendizado de Máquina, uma vez que se trata de um problema com dados incertos [Martin et al. 2005].

Para isso, a região selecionada para análise foi a Floresta Amazônica, mais especificamente, a cidade de Manaus (AM). Essa é considerada por especialistas como sendo um laboratório ideal para o estudo da influência da ação antrópica no clima e nos ecossistemas terrestres em florestas tropicais [Martin et al. 2017]. Assim, entre os anos de 2014 e 2015, foram realizadas coletas de dados meteorológicos e aerossóis, através de voos de baixa altitude. Esses fizeram parte do projeto *Green Ocean Amazon 2014/15* (GoAmazon 2014/15), organizado pelo *Atmospheric Radiation Measurement* (ARM), órgão ligado ao Governo dos Estados Unidos da América, em conjunto com instituições brasileiras [Martin et al. 2016]. A partir dessa coleta de dados foi possível obter as variáveis preditoras a serem utilizadas para a Modelagem de Distribuição de Espécies.

Já os dados utilizados como variável resposta relacionados à ocorrência das espécies foram coletados através de duas fontes: o Instituto Chico Mendes de Conservação da

Biodiversidade (ICMBio), que realiza o monitoramento da biodiversidade nacional e disponibiliza abertamente registro de ocorrência de espécies em todo o território do Brasil no Portal da Biodiversidade [ICMBio 2023]. Além do *Global Information Information Facility* (GBIF), um dos maiores repositórios de dados a respeito da ocorrência de espécies em todos os continentes [GBIF 2023]. Porém, para a região analisada no estudo de caso, agregando ambos os conjuntos de dados, obtém-se uma baixa quantidade de dados para uma mesma espécie. Esse fato pode dificultar a utilização de modelos de Aprendizado de Máquina [Almeida et al. 2021].

Para a construção do conjunto de dados para a Modelagem de Distribuição de Espécies, foram utilizados os dados resultantes do processo de interpolação espacial, que foi aplicado sobre os dados do projeto GoAmazon 2014/15 e disponibilizados por [Miyaji et al. 2021]. As variáveis disponibilizadas foram: a temperatura, as concentrações de ozônio (O_3), monóxido de carbono (CO), óxidos de nitrogênio (NO_X), metano (CH_4), dióxido de carbono (CO_2), isopreno, acetonitrila, a contagem numérica de partículas e a fração volumétrica de água (H_2O).

Para o mesmo período de análise e a região compreendida entre as mesmas coordenadas geográficas, foram obtidos os registros de ocorrência de espécies disponibilizados pelo ICMBio e pelo GBIF. Assim, as espécies que apresentaram maiores quantidades de ocorrência foram a *Coragyps atratus* (urubu de cabeça preta) e a *Tyrannus melancholicus* (suiriri), representando 54 e 50 registros distintos, respectivamente.

Utilizando a linguagem *Python*, foi construído o conjunto de dados bioclimáticos, realizando os devidos filtros, tratamentos e aplicando a operação de junção entre os conjuntos de dados, considerando as coordenadas geográficas e a data de registro das ocorrências. A espécie a ser analisada foi a com a maior frequência de ocorrências, a *Coragyps atratus*, por facilitar a aplicação de modelos de Distribuição de Espécies [Hernandez et al. 2006].

Devido ao natural desbalanceamento entre as classes da variável resposta e às incertezas associadas à marcação das classes negativas de ausência da espécie [Johnson et al. 2012], para amenizar o problema de Classificação Desbalanceada, optou-se pelo uso da técnica *Synthetic Minority Oversampling Technique* (SMOTE) [The Imbalanced-learn Developers 2021]. Trata-se de um método de rebalanceamento do conjunto de dados através do processo de reamostragem (*resampling*). Isso é feito através da criação de amostras sintéticas positivas, aumentando sua frequência. Para o estudo de caso, foi definido que o conjunto de dados reamostrado apresentasse a proporção de 1:3 entre amostras positivas e negativas. O tamanho do conjunto de dados era de 185305 linhas. No conjunto de dados original, a distribuição de amostras das diferentes classes era de 3,3% para a classe positiva (presença da espécie) e 96,7% para a classe negativa (ausência da espécie). Já no conjunto de dados rebalanceado através da técnica SMOTE, a distribuição era de 23% para a classe positiva e 77% para a classe negativa.

Em seguida, para selecionar as variáveis preditoras que seriam utilizadas no modelo de Distribuição de Espécies, foi realizada uma Análise de Correlação. Adotando o coeficiente de Pearson, analisou-se a relação linear entre as variáveis aos pares. Optou-se por retirar uma das variáveis dos pares com elevada correlação entre si, ou seja, com o coeficiente de Pearson com módulo maior ou igual a 80% [Mateo et al. 2013]. Assim,

evitou-se que o modelo incorporasse padrões aleatórios e o fenômeno de multicolinearidade [Pinaya and Corrêa 2014]. Foram retiradas três variáveis preditoras do conjunto de dados. Assim, as variáveis preditoras utilizadas foram: a temperatura máxima, a temperatura mínima, as concentrações de ozônio, monóxido de carbono, óxidos de nitrogênio, metano, isopreno, acetonitrila e a fração volumétrica de água.

O modelo de Classificação selecionado foi um dos mais comuns para a tarefa de Modelagem de Distribuição de Espécies: a Regressão Logística, por ser um modelo linear simples e um dos mais frequentes na literatura e com maior potencial [Hegel et al. 2010, Beery et al. 2021].

Definido o conjunto de dados e o modelo de Classificação, foi possível aplicar as técnicas de *Confident Learning* (CL) [Northcutt et al. 2021b], de acordo com o procedimento apresentado na Figura 1.

Como métricas de avaliação do modelo de Classificação, foram adotadas a acurácia, além da revocação, por se tratar de um problema de Classificação Desbalanceada. Ademais, também foi avaliada a *Area Under the Receiver Operating characteristic Curve* (ROC-AUC). As métricas foram avaliadas em conjunto de dados fora da amostra de treinamento (*out-of-sample*), a partir da técnica de Validação Cruzada com a divisão do conjunto de dados a partir do método *K-Fold* Estratificado, com $K = 5$, como recomendado para a aplicação da técnica *Confident Learning* [Northcutt et al. 2021b].

4. Resultados e Discussões

Aplicou-se a técnica *Confident Learning* (CL) a partir do modelo de Regressão Logística para o estudo de caso selecionado. Inicialmente, esse foi aplicado ao conjunto de dados original, modificado apenas com a incorporação de amostras sintéticas positivas através da técnica de reamostragem SMOTE. Através de seu uso, pode-se aprimorar a capacidade preditiva do classificador principalmente em relação à classe minoritária. Entretanto, por conta da criação de amostras sintéticas, pode-se introduzir incertezas maiores ao conjunto de dados. Nesse sentido, a aplicação de técnicas, como *Confident Learning* (CL), mostra-se ainda mais relevante.

Para a otimização dos hiper parâmetros do modelo, foi utilizada a Validação Cruzada, considerando o parâmetro C. Esse corresponde ao inverso da regularização, que controla a robustez do modelo a pequenas variações dos dados, evitando a ocorrência de sobreajuste [James et al. 2013]. Após esse processo, determinou-se o valor ideal como sendo de $C = 1$.

Em seguida, aplicou-se uma Validação Cruzada com o método *5-Fold*, para se mensurar as métricas de avaliação definidas, obtendo uma acurácia média de 76,7%, uma revocação de 50,0% para a classe positiva (minoritária) e um ROC-AUC de 78,1%. Assim, nota-se que o modelo foi capaz de desenvolver uma boa capacidade preditiva, porém isso não se reflete quando se considera apenas a classe minoritária que, para a aplicação de Modelagem de Distribuição de Espécies possui uma importância maior.

Então, o modelo treinado foi utilizado para a aplicação da técnica *Confident Learning*. Para isso, foram determinados os valores dos limites de classificação, ou seja, a auto-confiança esperada para cada uma das classes t_j. Foram obtidos os valores de $t_0 = 79,9\%$ para a classe negativa e $t_1 = 33,1\%$ para a classe positiva. Tal fato indica

que o modelo possui uma auto-confiança menor para a classe minoritária.

Definidas as auto-confiança esperadas, foi possível construir a matriz conjunta $C_{\tilde{y},y^*}[i][j]$. Em seguida, os valores foram normalizados, obtendo a distribuição de probabilidade conjunta $Q_{\tilde{y},y^*}[i][j]$, apresentada na Tabela 1. Assim, nota-se que cerca de 24,2% do conjunto de dados apresenta rótulos incorretos.

$Q_{\tilde{y},y^*}$	$y^* = 0$	$y^* = 1$
$\tilde{y} = 0$	62,4%	16,9%
$\tilde{y} = 1$	7,3%	13,4%

Tabela 1. Distribuição de probabilidade conjunta $Q_{\tilde{y},y^*}[i][j]$

Identificadas as observações incorretas, foi promovida uma limpeza no conjunto de dados através da Poda (*Pruning*), retirando-as. Então, um novo modelo de Regressão Logística foi treinado a partir do conjunto de dados tratado. Com o novo modelo, novamente aplicou-se uma Validação Cruzada com o método *5-Fold*, para se obter as métricas de avaliação. Foi obtida uma acurácia média de 95,6%, uma revocação de 75,0% para a classe positiva (minoritária) e um ROC-AUC de 95,8%. Portanto, nota-se uma melhoria significativa na capacidade preditiva do modelo em relação a todas as métricas avaliadas. Em específico, esse aprimoramento foi superior para as métricas relacionadas à classe positiva (minoritária), com incremento de 25 p.p. ou 50% na revocação. A comparação entre as métricas de desempenho dos modelos é apresentada na Tabela 2.

Conjunto de Dados	Acurácia	Revocação	ROC-AUC
Original	76,7%	50,0%	78,1%
Tratado com CL	95,6%	75,0%	95,8%

Tabela 2. Comparação das métricas de classificação para modelos treinados em diferentes conjuntos de dados

As melhores métricas de avaliação com a aplicação de métodos de *Confident Learning* (CL) são justificadas pelo fato de que o método é capaz de promover uma limpeza no conjunto de dados. Isso é possível através da identificação das observações nas quais existe uma incerteza maior sobre seu rótulo. Entretanto, para tal existe uma dependência grande em relação à auto-confiança esperada para cada classe, que determina se uma observação é considerada incorreta. Por isso, é relevante para o método que o classificador base possua uma capacidade preditiva elevada - como ocorre no estudo de caso. Se essa condição não for satisfeita, ou seja, a capacidade preditiva for muito baixa, esse critério torna-se pouco restritivo, sendo identificadas muitas observações como incorretas, o que pode prejudicar a análise desejada.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi possível avaliar a aplicação de técnicas da área de Inteligência Artificial centrada em Dados (*Data-Centric Artificial Intelligence* - DCAI) para realizar a limpeza de dados, o tratamento das incertezas associadas a eles e aprimorar o desempenho de modelos de Classificação de Aprendizado de Máquina. Em específico, foi aplicada a técnica de *Confident Learning* (CL), com o objetivo de se identificar observações do

conjunto de dados original, nas quais muito provavelmente havia um erro em seus rótulos de classe.

Para isso, a técnica CL foi aplicada sobre um estudo de caso de um experimento de Modelagem de Distribuição de Espécies. Trata-se de uma tarefa, na qual existem diversas fontes de incertezas associadas aos rótulos da variável resposta (a presença ou ausência da espécie analisada), posto que para a construção do conjunto de dados são utilizados apenas registros de presença das espécies (*Presence-only Data*). O conjunto de dados adotado se referia à espécie *Coragyps atratus*, com variáveis preditoras relacionadas a dados meteorológicos e aerossóis.

Por conta do desbalanceamento entre as classes no conjunto de dados utilizado para modelagem, foi aplicada a técnica de *Synthetic Minority Oversampling Technique* (SMOTE) para reamostrar a classe positiva. Em seguida, foi possível treinar um modelo de Regressão Logística, calculando suas métricas de desempenho. Esse modelo foi utilizado para aplicar a técnica *Confident Learning* para identificar as observações com rótulos possivelmente incorretos e realizar a limpeza dos dados. Dessa forma, observou-se um aumento de 76,7% para 95,6% na Acurácia e de 78,1% para 95,8% no ROC-AUC. A melhoria mais significativa foi em relação à métrica associada à classe minoritária, com um incremento de 50% (25 p.p.) na Revocação. Portanto, conclui-se que a técnica *Confident Learning* apresenta um grande potencial para o tratamento de incertezas em conjuntos de dados para a tarefa de Classificação de Aprendizado de Máquina, gerando uma melhoria significativa em seu desempenho.

Dado os resultados obtidos, para trabalhos futuros sugere-se a aplicação da técnica *Confident Learning* para outros classificadores de Aprendizado de Máquina comumente aplicados para a Modelagem de Distribuição de Espécies, como *Random Forests*, *Support Vector Machines*, Redes Neurais Artificiais, entre outros [Beery et al. 2021]. Nesses casos, como a capacidade preditiva do modelo pode ser superior, o critério adotado para avaliar se uma observação possui rótulo incorreto pode se tornar mais preciso, sendo possível aprimorar os resultados obtidos. Ademais, também podem ser aplicadas outras técnicas da área de Inteligência Artificial centrada em Dados (*Data-Centric Artificial Intelligence* - DCAI) que busquem tratar as incertezas presentes nos conjuntos de dados.

Agradecimentos

Este trabalho foi possível devido ao apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através de uma bolsa do programa PIBIC (2020/21 - 1745), dos Projetos Temáticos da FAPESP "Ciclos de vida e nuvens de aerossóis na Amazônia"(2017/ 17047-0) e "Research Centre for Greenhouse Gas Innovation - RCG2I"(2020/15230-5) e dos pesquisadores do Grupo de Pesquisa em Big Data e Ciência dos Dados da EPUSP.

Referências

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 (2023). Gbif | global biodiversity information facility. https://www.gbif.org/. Acesso em: 2023-05-14.

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.

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.

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 (2023). Portal da biodiversidade do instituto chico mendes de conservação da biodiversidade. https://portaldabiodiversidade.icmbio.gov.br/portal/. Acesso em: 2023-05-14.

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.

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. 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.

The Imbalanced-learn Developers (2021). Imbalanced-learn documentation. `https://imbalanced-learn.org/stable/`. Acesso em: 14/05/2023.

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