Comparação de Classificadores de Aprendizado de Máquina para Modelagem de Distribuição de Espécies: um estudo de caso na Bacia Amazônica

Renato O. Miyaji[1], Felipe V. de Almeida[1], Pedro L. P. Corrêa[1], Luciana V. Rizzo[2], Alan Calheiros[3], Márcio Teixeira[4]

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

[2]Universidade Federal de São Paulo (UNIFESP)

[3]Instituto Nacional de Pesquisas Espaciais (INPE)

[4]Instituto de Física – Universidade de São Paulo (USP)

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

{lrizzo}@unifesp.br

{alan.calheiros}@inpe.br

{mjt}@if.usp.br

Abstract. *In Ecology, Species Distribution Modeling is commonly performed to analyze the influence of atmospheric and meteorological variables on the occurrence of species. In the last decades, Machine Learning Classifiers have been successfully applied for this task. Therefore, this article aimed to compare the feasibility of seven Machine Learning techniques for Species Distribution Modeling. These were applied for a study case of the occurrence of birds in the central region of the Amazon Basin, near Manaus (AM), using data collected by the GoAmazon 2014/15 project. The classifier with the best ROC-AUC was Gradient Boosting with 94 %. Maximum Entropy Model obtained the best Recall with 85 %. Random Forests presented a good performance for both metrics.*

Resumo. *Na Ecologia, a Modelagem de Distribuição de Espécies é utilizada para analisar a influência de variáveis atmosféricas e meteorológicas na ocorrência de espécies. Nas últimas décadas, Classificadores de Aprendizado de Máquina foram aplicados com sucesso. Dessa forma, este artigo buscou comparar sete técnicas de Aprendizado de Máquina para avaliar sua viabilidade. Essas foram aplicadas para um estudo de caso sobre pássaros na região central da Bacia Amazônica próxima a Manaus (AM), com dados do projeto GoAmazon 2014/15. O classificador com melhor ROC-AUC foi o Gradient Boosting com 94 %. O modelo de Máxima Entropia teve a melhor revocação (85 %). O modelo Florestas Aleatórias teve a melhor ponderação entre as métricas.*

1. Introdução

Na área da Ecologia, uma análise de grande relevância é a Modelagem de Distribuição de Espécies (*Species Distribution Models* – SDM), dado que ela permite analisar

a influência de variáveis atmosféricas e meteorológicas na ocorrência de espécies [Elith and Leathwick 2009]. Ademais, ela também possibilita determinar o nicho ecológico da espécie analisada, que é definido pelas faixas de valores das variáveis que tornam um habitat adequado para a ocorrência da espécie [Hutchinson 1991].

Nas últimas décadas, os modelos de Aprendizado de Máquina ganharam cada vez mais complexidade, obtendo desempenhos elevados. Dessa forma, esses passaram a ser aplicados para a Modelagem de Distribuição de Espécies com mais frequência na literatura por meio de técnicas, como a Regressão Logística, Árvore de Decisão e Florestas Aleatórias [Hegel et al. 2010]. Nesse período, na literatura foram localizadas mais de 6000 publicações com a aplicação de modelos de Aprendizado de Máquina para esse fim [Araujo et al. 2019]. Porém, alguns classificadores recentes que apresentaram bom potencial para outras aplicações na literatura, como o *Extreme Gradient Boosting* (XGBoost), não foram explorados de maneira aprofundada para a Modelagem de Distribuição de Espécies [Effrosynidis et al. 2020].

Nesse contexto, este trabalho buscou aplicar diferentes classificadores de Aprendizado de Máquina, de modo a avaliar sua viabilidade para a tarefa de Modelagem de Distribuição de Espécies em relação às suas principais métricas de desempenho, determinando as técnicas mais promissoras.

Para isso, selecionou-se um estudo de caso sobre a ocorrência de espécies de pássaros na região central da Bacia Amazônica, entre as cidades de Manaus (AM) e Manacapuru (AM). Ela é considerada por especialistas como um laboratório ideal para estudar a influência da ação antrópica em ecossistemas terrestes em uma floresta tropical [Martin et al. 2017]. Os dados de variáveis atmosféricas e meteorológicas foram coletados no período entre janeiro de 2014 e dezembro de 2015, durante o projeto GoAmazon 2014/15. Esse foi organizado pelo *Atmospheric Radiation Measurement* (ARM), órgão do Departamento de Energia dos Estados Unidos da América, além de instituições brasileiras, como a Universidade de São Paulo (USP), a Universidade do Estado do Amazonas (UEA) e o Instituto Nacional de Pesquisas Espaciais (INPE), e realizou uma coleta extensiva de dados na região amazônica através de estações de pesquisa terrestres e voos realizados por aeronaves. Já os dados referentes à ocorrência de espécies podem ser obtidos pelo Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio), que realiza o monitoramento da biodiversidade nacional e disponibiliza os dados através do Portal da Biodiversidade. Desse modo, com ambas as fontes de dados, pôde-se realizar a Modelagem de Distribuição de Espécies.

Por meio deste trabalho, espera-se fornecer aos pesquisadores da área subsídios para selecionar as melhores técnicas de Classificação de Aprendizado de Máquina para serem aplicadas em estudos de caso de Modelagem de Distribuição de Espécies, utilizando recursos de Inteligência Artificial.

2. Trabalhos Relacionados

Para se identificar o Estado da Arte sobre técnicas de Classificação de Aprendizado de Máquina para a Modelagem de Distribuição de Espécies, foi realizada uma revisão bibliográfica sobre duas bases indexadoras de trabalhos científicos: *Scopus* e *Web of Science*. Por meio dela, buscou-se responder às seguintes questões de pesquisa: ”Quais técnicas de Inteligência Artificial já foram aplicadas para Modelagem de Distribuição de Espécies?”e

”Quais técnicas apresentaram os melhores desempenhos?”. Para isso, adotou-se a seguinte *string* de busca: (”Species Distribution” OR ”SDM” OR ”Distribution”) AND ((”Artificial Intelligence” OR ”AI”) OR (”Machine Learning” OR ”ML”)). O período de busca foi recente, considerando os anos de 2021 e 2020.

Assim, foi analisado um total de 50 trabalhos, sendo 19 publicados em 2020 e 31 em 2021. Neles, observou-se as técnicas de Classificação aplicadas, obtendo mais de 15 diferentes, como por exemplo: o modelo de Florestas Aleatórias (*Random Forests* - RF), o de Máxima Entropia (*MaxEnt*), o baseado em *Boosting*, o de Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM), as Redes Neurais Artificiais (*Artificial Neural Networks* - ANN), os modelos lineares generalizados (*Generalized Linear Models* - GLM), os modelos aditivos generalizados (*Generalized Additive Models* - GAM), as árvores de decisão (*Decistion Trees* - DT), a Regressão Logística (*Logistic Regression* - LR), o *Extreme Gradient Boosting* (XGBoost), o modelo *k-Nearest Neighbors* (k-NN), entre outros.

Entre essas diferentes técnicas, as mais frequentemente aplicadas nos artigos analisados foram: a de Florestas Aleatórias, o modelo de Máxima Entropia, os modelos baseados em *Boosting*, como o *Gradient Boosting* e o *Extreme Gradient Boosting*, e as técnicas de Máquinas de Vetores de Suporte. As demais técnicas não se mostraram tão recorrentes. Para responder à segunda questão de pesquisa, foram analisados os artigos que comparavam pelo menos quatro técnicas diferentes de Classificação de Aprendizado de Máquina. Nesses trabalhos, as métricas utilizadas para a avaliação dos classificadores foram: a AUC-ROC (*Area Under the Receiver Operating characteristic Curve*), a Acurácia e a Revocação.

Em trabalhos, como de [Rahman et al. 2021], [Nurhussen et al. 2021], [Carter et al. 2021], [Fern et al. 2020] e [Amado et al. 2020], a técnica de Florestas Aleatórias (*Random Forests* - RF) foi capaz de atingir um valor de ROC-AUC mais elevado que as demais técnicas comparadas. Já o modelo de Máxima Entropia apresentou o melhor desempenho de ROC-AUC para [Ghareghan et al. 2020] e [Derville et al. 2018]. O modelo de *Gradient Boosting*, obteve melhor ROC-AUC no trabalho de [Georgian et al. 2021]. Assim, apesar da modelagem dos trabalhos ter sido realizada para espécies diferentes, notou-se um maior potencial das técnicas de Florestas Aleatórias e modelo de Máxima Entropia.

Cabe ressaltar que não foram encontrados na literatura recente artigos com proposta semelhante a este, que busca comparar mais de seis classificadores de Aprendizado de Máquina para a tarefa de Modelagem de Distribuição de Espécies com base em variáveis meteorológicas e atmosféricas, incluindo técnicas recentes e promissoras, como o *Extreme Gradient Boosting* e o Modelo de Máxima Entropia.

3. Metodologia

A metodologia adotada para o experimento de Modelagem de Distribuição de Espécies foi adaptada de [Pinaya and Corrêa 2014], sendo composta por cinco etapas: a construção da hipótese científica, a pré-análise dos dados, a modelagem dos dados, a predição e a validação da hipótese científica. Essa é apresentada na Figura 1. Todo o experimento foi realizado na linguagem *Python*, utilizando a aplicação *Jupyter Notebook*.

A hipótese científica analisada foi de que a variação da concentração das variáveis

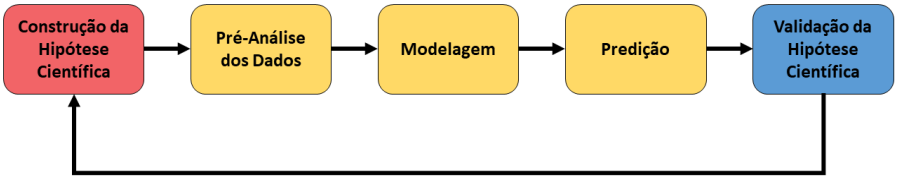

Figura 1. Metodologia para Modelagem de Distribuição de Espécies. Adaptada de [Pinaya and Corrêa 2014]

atmosféricas e meteorológicas, devido à ação antrópica, influencia a probabilidade de ocorrência de espécies na região central da Bacia Amazônica.

Para a construção do conjunto de dados sobre o qual a Modelagem de Distribuição de Espécies poderia ser realizada, foram coletados os dados atmosféricos e meteorológicos da região de interesse gerados por meio de interpolações espaciais realizadas e disponibilizadas no trabalho de [Miyaji et al. 2021]. Essas interpolações foram aplicadas de modo a expandir a área mapeada pelas aeronaves do projeto GoAmazon 2014/15, que realizaram mais de 35 voos de baixa altitude sobre a região durante as estações seca e chuvosa. Foram obtidas as seguintes variáveis: 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 e acetonitrila, a concentração numérica de partículas e a fração volumétrica de água (H_2O), sendo essas as possíveis variáveis preditoras. Também foi necessário coletar os dados de ocorrência de espécies sobre a mesma região de interesse. Esses foram obtidos a partir do repositório do Portal da Biodiversidade do ICMBio.

A partir dos dois conjuntos de dados, foi possível obter o conjunto de dados bioclimáticos sobre o qual poderia se desenvolver a Modelagem de Distribuição de Espécies. Isso foi feito por meio da operação de junção (*JOIN*), com base nas chaves de coordenadas geográficas (latitude e longitude) e data. O período considerado foi entre fevereiro e março de 2014 para a estação úmida e entre agosto e outubro de 2014 para a estação seca. Já em relação as coordenadas geográficas, foram consideradas: entre -3,632º e -2,813º de latitude e entre -60,831º e -59,937º de longitude.

Em seguida, foi necessário selecionar a espécie a ser analisada. Optou-se por uma espécie que possuísse a maior quantidade de pontos de ocorrência distintos e uma pequena dispersão espacial dentro da região analisada [Hernandez et al. 2006], selecionando a espécie *Coragyps atratus*, o urubu-de-cabeça-preta.

Na sequência, na etapa de Pré-análise dos Dados, selecionou-se as variáveis preditoras. Para isso, foi realizada uma Análise de Correlação, na qual foi analisado o coeficiente de correlação linear de Pearson entre cada variável do conjunto de dados bioclimáticos duas a duas. Foi retirada uma das variáveis preditoras que pertenciam aos pares altamente correlacionados com coeficiente de Pearson com módulo a partir de 80 %. Esse procedimento foi realizado com o objetivo de se evitar a multicolinearidade [Mateo et al. 2013]. A partir dessa análise, as variáveis de temperatura, concentração de dióxido de carbono e concentração numérica de partículas foram retiradas do conjunto de dados bioclimáticos.

Em seguida, na etapa de Modelagem, foram selecionados os classificadores a serem aplicados. Optou-se pelos apontados na revisão da literatura como de maior po-

tencial: a técnica de Florestas Aleatórias, o Modelo de Máxima Entropia, as técnicas baseadas em *Boosting* - *Gradient Boosting* e *XGBoost* e as Máquinas de Vetores de Suporte. Como base de comparação de desempenho, também foram aplicadas técnicas mais simples, como a Regressão Logística e as Árvores de Decisão. Dentre esses, existem modelos lineares, como a Regressão Logística e o Modelo de Máxima Entropia, e não lineares, como o de Árvores de Decisão e Máquinas de Vetores de Suporte. Os modelos de Florestas Aleatórias e de *Boosting* são baseados em Árvores de Decisão.

Quando se aborda a tarefa de Modelagem de Distribuição de Espécies como uma Classificação, um problema comum é o desbalanceamento entre as classes, uma vez que existem mais pontos de ausência da espécie (classe negativa) do que pontos de presença da espécie (classe positiva). Esse fato pode dificultar o ajuste de modelos de Classificação [Johnson et al. 2012]. Nesse tipo de problema, técnicas de rebalanceamento (*resampling*) do conjunto de dados podem ser úteis. Para isso, aplicou-se a técnica *Synthetic Minority Oversampling Technique* (SMOTE) para criação de amostras sintéticas positivas, aumentando sua frequência [The Imbalanced-learn Developers 2021]. Definiu-se que o conjunto de dados rebalanceado deveria possuir a proporção de 1:3 para amostras positivas e negativas.

Após a etapa de Predição, para a comparação do desempenho das técnicas, uma métrica avaliada foi a ROC-AUC. Além disso, também foram avaliadas métricas de classificação. Uma delas foi a Acurária, que é capaz de medir a taxa de acertos do modelos. Também foi avaliada outra métrica que se refere especificamente ao desempenho do modelo para a classe positiva. Essa deve ser considerada principalmente em problemas de Classificação Desbalanceada. A métrica avaliada foi a Revocação, que é definida pela razão entre os Verdadeiros Positivos e a somatória de Verdadeiros Positivos e Falsos Negativos, apresentada na equação 1.

$$Revocacao = \frac{VP}{VP + FN} \tag{1}$$

Para a validação dos modelos comparados, separou-se um conjunto de dados de *Hold-out* de maneira aleatória, utilizando a proporção 70 % e 30 %. Também foi adotada um outra forma de validação para que se pudesse otimizar os valores dos hiper parâmetros de cada um dos modelos, de modo a se obter as configurações com os melhores desempenhos, considerando as métricas avaliadas. Isso foi feito por meio de uma Validação Cruzada com método *K-fold* com $K = 5$. Para o modelo de Regressão Logística, foi otimizado o hiper parâmetro de regularização. Para os modelos baseados em Árvores de Decisão, os relacionados à construção da árvore, como profundidade máxima, amostra mínima por folhas, entre outros. Para o modelo SVM, o tipo de *kernel* e regularizações.

Então, na etapa de Validação da Hipótese Científica, os resultados obtidos pelos melhores modelos testados podem ser avaliados por pesquisadores especialistas na área e pela interpretação dos modelos [Pinaya and Corrêa 2014].

4. Resultados e Discussões

A partir da metodologia apresentada, os classificadores selecionados foram ajustados e aplicados para o estudo de caso. Na Tabela 1 são apresentadas as métricas de desempenho dos diferentes modelos no conjunto de *Hold-out*.

Classificador	Acurácia	Revocação	ROC-AUC
RL	96 %	6 %	67 %
DT	99 %	50 %	83 %
RF	98 %	62 %	90 %
SVM	94 %	62 %	85 %
GB	99 %	38 %	94 %
XGBoost	99 %	38 %	89 %
MaxEnt	75 %	85 %	80 %

Tabela 1. Métricas de classificação para os modelos avaliados

Analisando os resultados apresentados na Tabela 1, percebe-se que os classificadores obtiveram uma Acurácia elevada, exceto o de Máxima Entropia. Porém, quando se analisa as demais métricas de Classificação, o desempenho é diferente.

Para o classificador de Regressão Logística (RL), a Acurácia é elevada de 96 %, porém a Revocação é apenas de 6 %. Isso ocorre devido ao problema de Classificação Desbalanceada: apesar do rebalanceamento aplicado no conjunto de dados, o modelo de Regressão Logística obteve o viés de classificar a maior parte das amostras como a classe majoritária (negativa), de modo a maximizar a Acurácia. Porém, ao mesmo tempo, não foi capaz de identificar as amostras referentes à classe minoritária (positiva). Assim, a quantidade de Verdadeiros Positivos (VP) é pequena e as quantidades de Falsos Negativos (FN) e Falsos Positivos (FP) são grandes, tornando a métrica de Revocação baixa. A Regressão Logística obteve um ROC-AUC de 67 %, sendo a menor entre todos os modelos avaliados.

O classificador de Árvore de Decisão (DT), por ser um modelo com maior capacidade de identificar padrões complexos dentro do conjunto de dados, como não linearidades, obteve uma Revocação bem superior ao de Regressão Logística: de 50 %. Além de um ROC-AUC 16 pontos percentuais maior: de 83 %.

Os outros modelos baseados em Árvores de Decisão, como as Florestas Aleatórias (RF), *Gradient Boosting* (GB) e *Extreme Gradient Boosting* (XGBoost), apresentaram desempenhos ainda melhores em relação ao classificador de Árvore de Decisão (DT). Em termos de ROC-AUC, o melhor classificador foi de *Gradient Boosting* com 94 %, Florestas Aleatórias com 90 % e *Extreme Gradient Boosting* com 89 %. Já em termos de Revocação, o modelo de Florestas Aleatórias apresentou a melhor capacidade de identificar as classes minoritárias (positivas), obtendo uma Revocação de 62 %. O desempenho dos modelos de *Boosting* foi inferior, com 38 % de Revocação. Tal fato indica que os modelos foram capazes de aumentar a quantidade de Verdadeiros Positivos (VP) e reduzir os Falsos Negativos (FN), levando a uma alta Revocação.

Para o modelo de Máquinas de Vetores de Suporte (SVM), o desempenho foi semelhante aos de RF, SVM e GB, mantendo um patamar elevado de Acurácia de 94 %, uma Revocação tão alta quanto ao modelo de Florestas Aleatórias de 62 % e um ROC-AUC um pouco inferior: de 85 %.

Por fim, o Modelo de Máxima Entropia (MaxEnt) apresentou uma Acurácia inferior em relação aos demais classificadores de 75 %, porém obteve a melhor capacidade de

reduzir os Falsos Negativos (FN) dentre todas as técnicas avaliadas, com uma Revocação de 85 %. O ROC-AUC foi semelhante aos classificadores de Árvore de Decisão e de Máquinas de Vetores de Suporte (SVM), com 80 %.

Dessa forma, considerando as métricas de Classificação avaliadas, o melhor classificador para ROC-AUC foi o *Gradient Boosting* com 94 %, apesar de apresentar Revocação inferior. Já considerando a métrica de Revocação, o melhor foi o de Máxima Entropia com 85 %, apesar de apresentar ROC-AUC inferior. O modelo que apresentou os melhores ROC-AUC e Revocação ao mesmo tempo foi o de Florestas Aleatórias. Nota-se que essas três técnicas foram justamente as mais aplicadas na literatura.

Assim como no trabalho de [Georgian et al. 2021], o modelo que apresentou o melhor desempenho em termos de ROC-AUC foi o baseado em *Boosting*, quando comparado com outras técnicas (Máxima Entropia, Florestas Aleatórias, entre outros). Apesar de não apresentar o melhor ROC-AUC, o modelo de Florestas Aleatórias mostrou-se com uma boa capacidade preditiva tanto para Acurácia, quanto Revocação e ROC-AUC, como apontado nos trabalhos de [Rahman et al. 2021], [Nurhussen et al. 2021], [Carter et al. 2021], [Fern et al. 2020] e [Amado et al. 2020].

Para validar a hipótese científica estabelecida de que a variação nas concentrações das variáveis preditoras influencia a probabilidade de ocorrência de espécies na região de interesse, optou-se pelo modelo que apresentou a melhor ponderação entre ROC-AUC e Revocação: o de Florestas Aleatórias.

Para ele, o recurso relacionado à interpretação do modelo que foi utilizado foi a medida da importância de cada variável preditora devido ao decréscimo médio de impuridade [Breiman 2001], apresentada na Figura 2. Nela, observa-se que as variáveis mais relevantes para o modelo de Florestas Aleatórias foram: as concentrações de isopreno (C_8H_5), óxidos de nitrogênio (NO_X), a temperatura mínima e as concentrações de acetonitrila (CH_3CN) e metano (CH_4).

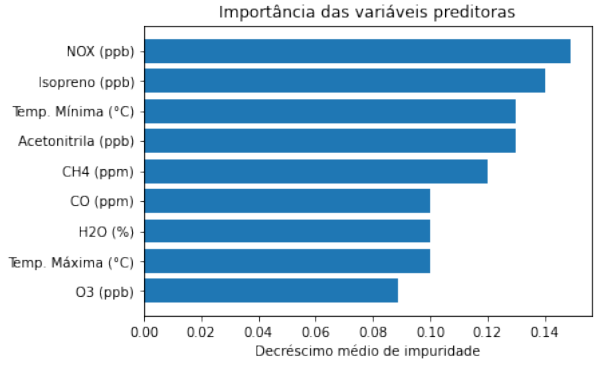

Figura 2. Importância de cada variável do modelo de Florestas Aleatórias para a espécie *Coragyps atratus*

Ademais, também foram desenhados os Gráficos de Dependência Parcial para o modelo de Florestas Aleatórias. Esses representam a dependência da variável-resposta que se está prevendo, em função da variação de cada uma das variáveis preditoras e podem ser vistos na Figura 3, na qual estão apresentados seus valores normalizados. Através da análise, percebe-se que a principal variável com efeito negativo na probabilidade de

ocorrência de *Coragyps atratus* é a concentração de isopreno, posto que a probabilidade é nula a partir de 1,4 ppb de concentração.

Dessa forma, foi possível observar que principalmente as variáveis de isopreno (C_8H_5), óxidos de nitrogênio (NO_X), a temperatura mínima e as concentrações de acetonitrila (CH_3CN) e monóxido de carbono (CO) possuem influência na probabilidade de ocorrência da espécie *Coragyps atratus*.

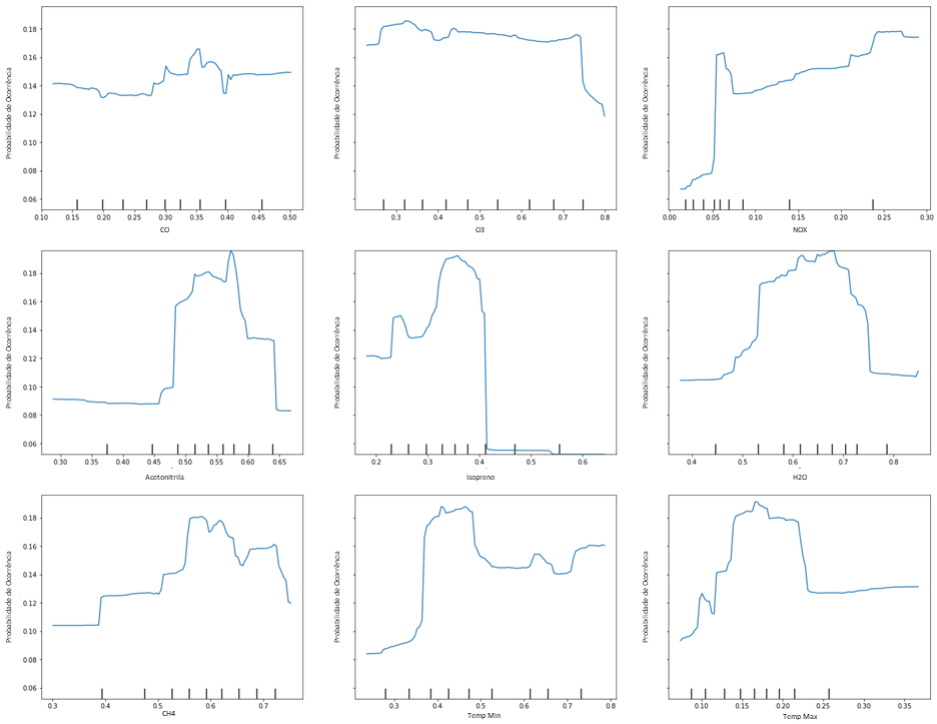

Figura 3. Gráficos de Dependência Parcial do modelo de Florestas Aleatórias para probabilidade de ocorrrência da espécie *Coragyps atratus* em função das variáveis preditoras (valores normalizados)

5. Conclusão e Trabalhos Futuros

Com os resultados obtidos neste trabalho, conclui-se que foi possível avaliar a viabilidade da aplicação de técnicas de Classificação de Aprendizado de Máquina para a Modelagem de Distribuição de Espécies. Para isso, realizou-se uma revisão bibliográfica, na qual foi possível identificar as principais técnicas utilizadas na literatura recente com maior potencial. Então, através da coleta de dados atmosféricos e de aerossóis do projeto GoAmazon 2014/15 e de dados de ocorrência de espécies, gerou-se um conjunto de dados sobre o qual o estudo de caso poderia ser desenvolvido.

Foram comparadas as técnicas de Florestas Aleatórias, o Modelo de Máxima Entropia, as técnicas baseadas em *Boosting* - como o *Gradient Boosting* e *XGBoost* - as Máquinas de Vetores de Suporte, a Regressão Logística e as Árvores de Decisão. Para o estudo de caso, o melhor classificador para ROC-AUC foi o *Gradient Boosting* com 94 %, apesar de apresentar Revocação inferior. Já considerando a métrica de Revocação, o melhor foi o de Máxima Entropia com 85 %, apesar de apresentar ROC-AUC inferior. O modelo que apresentou os melhores ROC-AUC e Revocação ao mesmo tempo foi o

de Florestas Aleatórias. Os resultados obtidos pelo modelo de Florestas Aleatórias foram analisados para se validar a hipótese científica estabelecida.

Para se aprimorar ainda mais os resultados obtidos, podem ser aplicadas outras técnicas de rebalanceamento do conjunto de dados, de modo a reduzir o efeito do desbalanceamento entre classes da variável-resposta. Também podem ser testados outros classificadores, como as Redes Neurais Artificiais. A inclusão de mais dados climáticos como variáveis preditivas também pode aprimorar o desempenho dos modelos. Já para tornar os modelos mais interpretáveis, podem ser aplicados outros recursos de Inteligência Artificial Explicável. Também pode ser realizada uma revisão da literatura para um período maior, de modo a se identificar outras possíveis abordagens para o problema.

Agradecimentos

O presente trabalho foi possível devido a disponibilidade dos dados nos repositórios do *GOAmazon* e ICMBio e aos 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 aos pesquisadores do Grupo de Pesquisa em Big Data e Ciência dos Dados da EPUSP.

Referências

Amado, M. E. V., Grütter, R., Fischer, C., Suter, S., and Bernstein, A. (2020). Free-ranging wild boar (sus scrofa) in switzerland: Casual observations and model-based projections during open and closed season for hunting. *Schweiz Arch Tierheilkd*, 162(6):365–376.

Araujo, M. B., Anderson, R. P., Barbosa, M. A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., and Rhabek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Carter, S., van Rees, C. B., Hand, B. K., Muhlfeld, C. C., Luikart, G., and Kimball, J. S. (2021). Testing a generalizable machine learning workflow for aquatic invasive species on rainbow trout (oncorhynchus mykiss) in northwest montana. *Frontiers in Big Data*, 4.

Derville, S., Torres, L. G., Iovan, C., and Garrigue, C. (2018). Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches. *Diversity and Distributions*, 24:1657–1673.

Effrosynidis, D., Tsikliras, A., Arampatzis, A., and Sylaios, G. (2020). Species distribution modelling via feature engineering and machine learning for pelagic fishes in the mediterranean sea. *Applied Sciences*, 10(24).

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *The Annual Review of Ecology, Evolution and Systematics*, 40:677–697.

Fern, R. R., Morrison, M. L., Grant, W. E., Wang, H., and Campbell, T. A. (2020). Modeling the influence of livestock grazing pressure on grassland bird distributions. *Ecological Processes*, 9(42).

Georgian, S., Morgan, L., and Wagner, D. (2021). The modeled distribution of corals and sponges surrounding the salas y gómez and nazca ridges with implications for high seas conservation. *Peer J*, 9.

Ghareghan, F., Ghanbarian, G., Pourghasemi, H. R., and Safaeian, R. (2020). Prediction of habitat suitability of morina persica l. species using artificial intelligence techniques. *Ecological Indicators*, 112.

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.

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

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

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.

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., Almeida, F. V., Bauer, L. O., Ferrari, V., Corrêa, P. L. P., Rizzo, L. V., and Prakash, G. (2021). Spatial interpolation of air pollutant and meteorological variables in central amazonia. *Data*, 6(12).

Nurhussen, A., Atzberger, C., and Zewdia, W. (2021). Species distribution modelling performance and its implication for sentinel-2-based prediction of invasive prosopis juliflora in lower awash river basin, ethiopia. *Ecological Processes*, 10(18).

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.

Rahman, M. S., Pietong, C., Zafar, S., Ekalasananan, T., Paul, R. E., Haque, U., Rocklöv, J., and Overgaard, H. J. (2021). Mapping the spatial distribution of the dengue vector aedes aegypti and predicting its abundance in northeastern thailand using machine-learning approach. *One Health*, 13.

The Imbalanced-learn Developers (2021). Imbalanced-learn documentation. `https://imbalanced-learn.org/stable/`. Acesso em: 18/08/2022.