Explorando técnicas de aprendizado de máquina para aprimoramento da previsão de geadas no sul e sudeste do Brasil

José Roberto Motta Garcia

Instituto Nacional de Pesquisas Espaciais (INPE)
Cachoeira Paulista – SP – Brasil

roberto.garcia@inpe.br

Abstract. *The CPTEC/INPE operational frost forecast is based on an index established from the weighted variation of meteorological variables from a weather forecasting model, which is compared to the operationally predicted values. Constant improvement of these forecasts is necessary to mitigate the economic and social damage caused by frosts. The objective of this work is to identify a methodology based on machine learning that surpasses the current prediction system. Therefore, this exploratory research applied different combinations between three algorithms, varying their parameterizations and the set of predictor variables from meteorological station data in southern and southeastern Brazil. Preliminary tests on 24-hour forecasts show that performance surpasses the current method, signaling that the search for the most suitable methodology should be continued.*

Resumo. *A previsão operacional de geadas do CPTEC/INPE é baseada em um índice estabelecido a partir da variação ponderada das variáveis meteorológicas de um modelo de previsão de tempo, que é comparado aos valores previstos operacionalmente. Um constante aprimoramento dessas previsões visando mitigar danos econômicos e sociais decorrentes de geadas se faz necessária. O propósito deste trabalho é identificar uma metodologia baseada em aprendizado de máquina que supere o desempenho do sistema de previsão atual. Para tanto, esta pesquisa exploratória aplicou diferentes combinações entre três algoritmos, variando suas parametrizações e o conjunto de variáveis preditoras de dados de estações meteorológicas do sul e sudeste do Brasil. Testes preliminares em previsões de 24h mostram que o desempenho supera o método atual, sinalizando que o processo de busca da metodologia mais adequada deve ser continuada.*

1. Introdução

Geadas causam impacto na agricultura e refletem na economia (Aguiar et al., 2004). Tecnicamente, uma geada é dada como ocorrida quando há formação de cristais de gelo nas superfícies de plantas e objetos expostos ao relento (Pereira et al., 2002). A definição de geada neste trabalho segue o critério estabelecido do Índice de Geadas do CPTEC de Rozante et al. (2020) (IG-CPTEC, de agora em diante) que considera a temperatura de 6°C como limiar para a ocorrência ou não de geadas. Como parte do protocolo de proteção de lavouras, prever ocorrências de geadas se apresenta como um fator de grande importância para minimizar esses impactos.

O IG-CPTEC é estabelecido a partir de pesos decorrentes de cálculos sobre as médias e desvios padrão das principais variáveis meteorológicas que favorecem ou se contrapõem à ocorrência de geadas, apenas para os casos em que elas ocorrem, conforme definido pela metodologia. Essas variáveis são extraídas do modelo regional Eta (Mesinger et al., 1988; Black, 1994; Chou et al., 2002) que são submetidas a um processo de calibração. Este processo sinalizou que a temperatura tem a maior contribuição para a ocorrência de geadas, seguida pela pressão e ventos. As demais variáveis foram ajustadas para que a soma de pesos fosse igual a 1.

A metodologia de predição aqui proposta usa técnicas de aprendizado de máquina (do Inglês Machine Learning, ML daqui em diante) para construir modelos estatísticos a partir de dados históricos visando prever geadas por classificação binária, assumindo 1 para a ocorrência ou 0 para a não ocorrência. Como este trabalho visa melhorar o IG-CPTEC foram mantidas todas as premissas nele contidas e dados por ele utilizados. Uma breve revisão bibliográfica mostra que vários algoritmos de aprendizado de máquina têm sido usados para esta tarefa, como em Naing e Htike (2015) e em Robinson e Mort (1997).

Como exploração foram usados os algoritmos Florestas Aleatórias (RF) de Liaw e Wiener (2002), Redes Neurais Artificiais (ANN) de Haykin (1999) e uma variação de Gradient Boosting Machine (GBM) de Friedman (2001) adaptados para classificação binária. Seus hiperparâmetros, além de variações do conjunto de preditores utilizados para treinamentos dos modelos também foram explorados. O resultado final mostra que a modelagem é bastante promissora pois supera o IG-CPTEC nas previsões de 24h em todos os experimentos.

2. Metodologia

2.1. Dados

Neste trabalho foram avaliadas apenas as previsões de 24h do modelo regional Eta. As variáveis utilizadas como preditoras são: latitude e longitude da estação meteorológica (LON, LAT); temperatura em 2m (TP2M), pressão ao nível médio do mar (PNMM), magnitude do vento em 10m (V10M), nebulosidade (NEBUL) e umidade relativa (UMID) previstas, além da topografia (altitude) da localização da estação meteorológica (TOPO). Além disso há foi atribuído 1 quando houve geada e 0 quando não houve geada na localidade e data válida da previsão, dado que foi reportado pelo Instituto Nacional de Meteorologia (INMET). Os dados de treinamento compreendem 5 anos de

previsão (2012 a 2016), dos meses de maio a setembro, totalizando 244.096 casos de treino. Para os dados de avaliação (teste) foram usadas as previsões do ano de 2017 para os mesmos meses, totalizando 47.112 casos. A distribuição geográfica das estações meteorológicas é mostrada na Figura 1, que mostra também as regiões utilizadas para avaliar o IG-CPTEC.

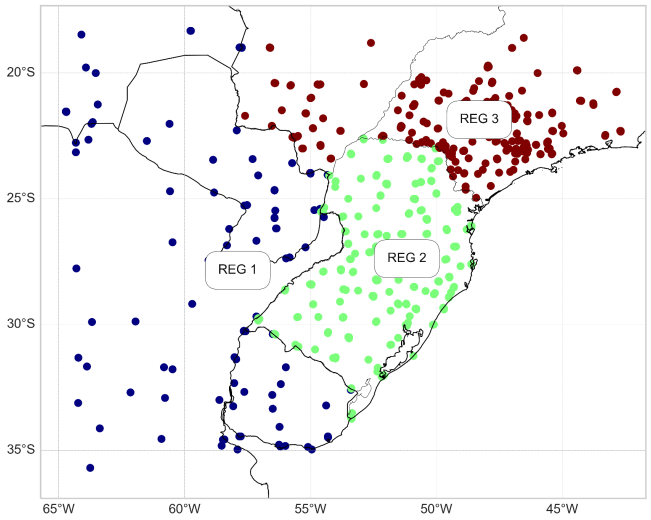


Figura 1. Domínio espacial

2.2. Métricas

As métricas aplicadas na avaliação são baseadas na matriz de confusão para classificação binária, que mostra os acertos e erros das estimativas (previsões) perante a real ocorrência ou não do evento (a geada). A partir dessa matriz de confusão podem ser calculados vários índices estatísticos (Stephenson, 2000 e Wilks, 2011) mas, para facilitar a visualização do desempenho, foi utilizado o diagrama de desempenho (em inglês, performance diagram) (Roebber, 2009), que sintetiza algumas desses índices num único diagrama, seguindo a metodologia praticada em IG-CPTEC.

2.3. Pesquisa exploratória: algoritmos de classificação binária e hiperparâmetros

Um objetivo secundário deste trabalho é descobrir a configuração ótima dos hiperparâmetros das funções que implementam os algoritmos. Foi feita uma exploração exaustiva *ad-hoc* pois podem influenciar positiva ou negativamente o modelo quando aplicadas a diferentes problemas (Yang L. e Shami, A. (2020). Toda a modelagem foi feita em Python (Python Org, 2021) e os algoritmos, explorados através do pacote Scikit-learn (Pedregosa et al., 2011), conforme descrito abaixo:

- **Random Forest (RF):** n_estimators (100, 200, 500, 1000, 1500, e 2000); max_leaf_nodes (None, 5, 25, 50, 100, 250, 500, 600, 700, 800, 1000, 1500, e 2000), e class_weight ("*balanced*" e "*balanced_subsample*").
- **XGBoost:** n_estimators (100, 200, 500, 1000, 1500, e 2000); max_depth (6, 8, e 10), e learning_rate (0.1, 0.05, 0.01, 0.005, e 0.001)

- **Redes Neurais Artificiais (ANN):** hidden_layer_sizes ([6,3], [6,6], [6,12], [12,6], e [12,6,3]); activation "relu" e "logistic"; solver ("sgd" e "lbfgs"), e learning_rate ("adaptive" e "invscaling")

2.4. Exploração dos preditores

Embora o IG-CPTEC tenha se baseado nas variáveis: TP2M, PNMM, V10M, UMID, e NEBUL para ser calculado, este trabalho também explorou variações deste conjunto de preditores, com o acréscimo de TOPO, visando verificar a influência de cada um deles sobre o resultado final dos modelos. Dessa maneira, houve a exploração do uso de todos estes preditores como entrada dos algoritmos até que restasse apenas um. O método de eliminação se baseou nos resultados obtidos em cada treinamento onde sempre o preditor menos influente era eliminado. Como mostrado na Figura 2.

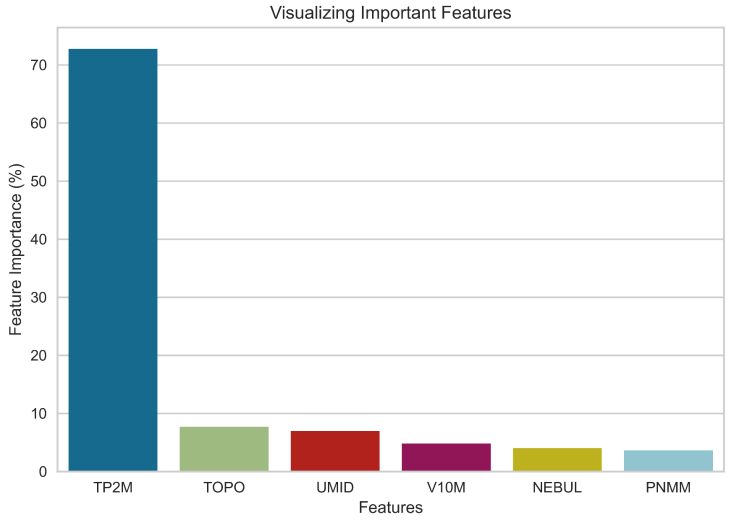


Figura 2. Exemplo da influência de cada preditor nos modelos estatísticos

2.5. Resultados da exploração de algoritmos e preditores

Para facilitar a visualização dos resultados, assume-se que os preditores TP2M, PNMM, V10M, NEBUL, UMID e TOPO são representados pelas letras a, b, c, d, e, e f. As Tabelas 1, 2 e 3 mostram os três melhores resultados da exploração dos algoritmos selecionados, dos seus hiperparâmetros e dos preditores. Há também o tempo gasto em segundos em cada treinamento (SEG).

Tabela 1. Rank dos três primeiros resultados da exploração via RF

PREDITORES	N_ESTIM	MAX_NODES	CLASS_WEIGHT	HR	SEG
a+b+c+d+e+f	1500	None	bal_subsample	0.963195	222
a+b+c+d+e+f	500	None	bal_subsample	0.963195	86

a+b+c+d+e+f	200	None	bal_subsample	0.963174	39

Tabela 2. Rank dos três primeiros resultados da exploração via XGBoost

PREDITORES	N_ESTIM	MAX_DEPTH	LEARN_RATE	HR	SEG
a+b+c+d+e+f	1500	8	0.010	0.967419	737
a+b+c+d+e+f	2000	8	0.010	0.967398	881
a+b+c+d+e+f	2000	10	0.005	0.967185	1426

Conforme pode ser visto nas Tabelas 1 e 2, os melhores resultados, ordenados de forma decrescente pelo Hit Rate (HR), foram conseguidos usando todos os preditores. Dessa maneira os testes utilizando redes neurais foram feitos sem variação nos preditores, uma vez que havia fortes indícios de que a utilização de todos os preditores era a melhor opção. A Tabela 3 mostra as configurações dos hiperparâmetros utilizados nos três melhores resultados utilizando ANN.

Tabela 3. Rank dos três primeiros resultados da exploração via ANN

HIDDEN_LAYERS	ACTIV	SOLVER	LEARN_RATE	HR	SEG
(6, 12)	relu	lbfgs	adaptive	0.962664	667
(6, 12)	relu	lbfgs	invscaling	0.962664	507
(6, 12)	logistic	lbfgs	invscaling	0.962643	521

A melhor dentre todas as variações exploradas foi conseguida via XGBoosting com a seguinte configuração de hiperparâmetros: N_ESTIMATORS=1500, MAX_DEPTH=8 e LEARNING_RATE=0.01, que obteve o melhor Hit Rate (HR) de 0.967419 levando 737s para ser treinado. Portanto esta configuração foi considerada para realizar as avaliações finais. Além disso, como já informado, os melhores resultados foram conseguidos usando todos os preditores. Embora os resultados sejam muito similares, XGBoost foi o algoritmo que obteve o melhor desempenho, medido pelo melhor Hit Rate (HR) e foi a opção considerada para as avaliações finais.

2.6. Experimentos

Visando descobrir a melhor maneira de agrupar os dados de entrada dos algoritmos, foram realizados quatro experimentos. Os resultados são apresentados por região, como feito em IG-CPTEC, e também forma generalizada do modelo para todo o domínio. Os seguintes experimentos foram conduzidos:

- EXP 1 - ONE-FITS-ALL: Um modelo XGBoost genérico, aplicado a todos os dados de teste. Neste experimento, todos os dados de treinamento foram submetidos para construir o modelo, sem qualquer distinção ou classificação, resultando em apenas um modelo. Visa verificar se existe um comportamento

generalizado da atmosfera quando há ou não há geada e se este comportamento foi capturado pelo modelo. A Figura 3a mostra a distribuição espacial considerada.

- EXP 2 - BY-REGION: Um modelo XGBoost para cada uma das três regiões definidas em IG-CPTEC, aplicado de forma distinta a todos os dados de teste, conforme a região que o dado se encontra, verificado pelo par LON+LAT. Neste experimento, os dados de treinamento e teste foram separados pela região que se encontram e, em seguida, submetidos para construir os modelos, resultando em três modelos estatísticos. Visa verificar se existe um comportamento específico da atmosfera para cada região definida em IG-CPTEC e se este comportamento foi capturado pelo modelo. A Figura 1 mostra as regiões consideradas.

Visando encontrar similaridades "comportamentais" entre os dados de treinamento, foram utilizadas duas abordagens de agrupamento dos preditores: por semelhança climática (CLUST-CLIM) e por semelhança entre localidades (CLUST-LOC). Foi utilizado o algoritmo KMeans do pacote Scikit-learn e a função z-score do pacote Scipy para normalizar os dados. A determinação da quantidade ideal de *clusteres* foi feita pelo método *elbow* (cotovelo), apresentado por Thorndike (1953). O processo resultou numa divisão dos dados em oito grupos, tanto para CLUST-LOC como para CLUST-CLIM, que foram associados aos dados de entrada. Conforme a Figura 3b, a linha azul refere-se a curva dos resultados do grau de distorção (escala azul à esquerda) obtidos pelo modelo para cada valor de K; o ponto de inflexão está anotado com uma linha tracejada (e também informado no canto superior direito do gráfico); a linha verde refere-se a escala de tempo gasto para resolver o modelo (escala verde à direita).

- EXP 3 - CLUST-CLIM: Um modelo XGBoost para cada um dos oito clusteres de comportamento climático similar, considerando todas as variáveis, exceto LON e LAT. Cada caso de treinamento foi associado a um dos oito grupos estimados para CLUST-CLIM e, em seguida, treinados, resultando em oito modelos estatísticos. Para testar a modelagem, cada caso de teste foi comparado aos centróides dos oito clusteres criados e foi usado o modelo estatístico referente ao centróide mais semelhante ao dado, medido por distância euclidiana. Visa verificar se a abordagem consegue distinguir comportamentos climáticos considerando cada medição da atmosfera. Não há representação espacial desse agrupamento.
- EXP 4 - CLUST-LOC: Um modelo XGBoost para cada uma das oito localidades de comportamento climático similar, considerando a média das variáveis para cada par único de LON+LAT, o que resultou em 593 diferentes pares LON+LAT. Cada caso de treinamento foi associado a um dos oito grupos estimados para CLUST-LOC, e em seguida treinados, resultando em oito modelos estatísticos. Para testar a modelagem, cada caso de teste foi comparado aos pares únicos de LON+LAT do *dataset* de treinamento e foi usado o modelo estatístico referente ao par de LON+LAT mais próximo do caso de teste, medido por distância euclidiana. Visa verificar se a abordagem consegue distinguir zonas climáticas de acordo com o

comportamento médio das estações meteorológicas. A Figura 3c mostra a clusterização das zonas climáticas de CLUST-LOC.

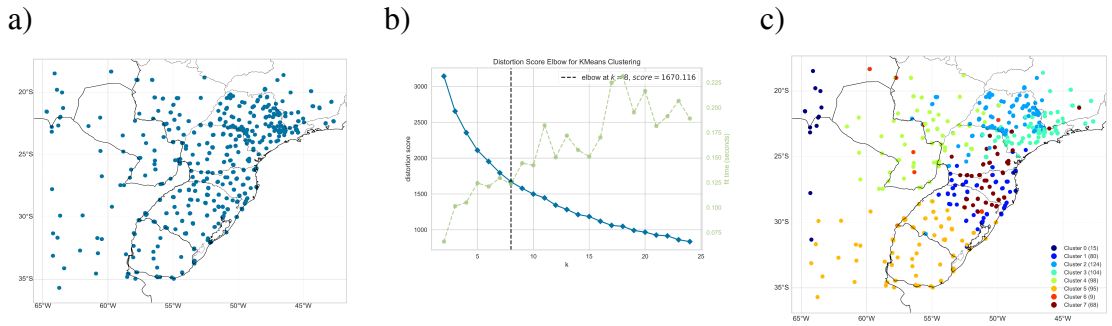


Figura 3. Domínio espacial do experimento ONE-FITS-ALL (a); exemplo de saída do método "Elbow" (b), e clusterização resultante da similaridade encontrada entre os pares LON+LAT únicos em CLUST-LOC (c).

3. Resultados

O desempenho comparativo entre IG-CPTEC e os experimentos deste trabalho pode ser analisado comparando as Figuras 4 e 5, referentes aos diagramas de desempenho das metodologias. Esse diagrama facilita a análise dos resultados, pois resume várias métricas obtidas a partir de estimativas dicotômicas (sim ou não) como: POD (Probability of Detection), FBias (Frequency Bias), CSI (Critical Success Index), FAR (False Alarm Ratio) e SR (Success Ratio). As métricas POD, CSI e SR são medidas de sucesso no desempenho então quanto maior seu valor melhor. POD pode ser visualizado no eixo vertical esquerdo, CSI é representado pelas curvas que convergem para o canto superior direito e sua escala é posicionada no eixo vertical direito. SR é representado pelo posicionamento do marcador dentro do gráfico. FBias é representado pelas linhas radiais de origem em (0,0) e indicam a frequência que as estimativas são sub ou superestimadas. Como FAR é uma medida de erro, o seu valor é subtraído de um para ser posicionado no eixo horizontal inferior, fazendo com que quanto mais à direita esteja melhor seja. Em resumo, quanto mais próximo o marcador da estimativa estiver posicionada do canto superior direito e sobre a diagonal principal melhor ela é.

A Figura 4 mostra os três Diagramas de Desempenho mostrados em IG-CPTEC, para a Região 1 (4.b), Região 2 (4.d) e Região 3 (4.f). Em cada um deles o círculo aberto menor representa as previsões de 24h, nas quais este trabalho foi baseado. Os outros círculos abertos maiores se referem a períodos de previsão até 120h, não considerados neste trabalho. Já os círculos fechados em cinza referem-se ao desempenho do modelo Eta, caso fosse considerado apenas a sua temperatura para prever a ocorrência ou não de geadas. A interpretação é que o IG-CPTEC conseguiu melhorar esta previsão através de seu índice.

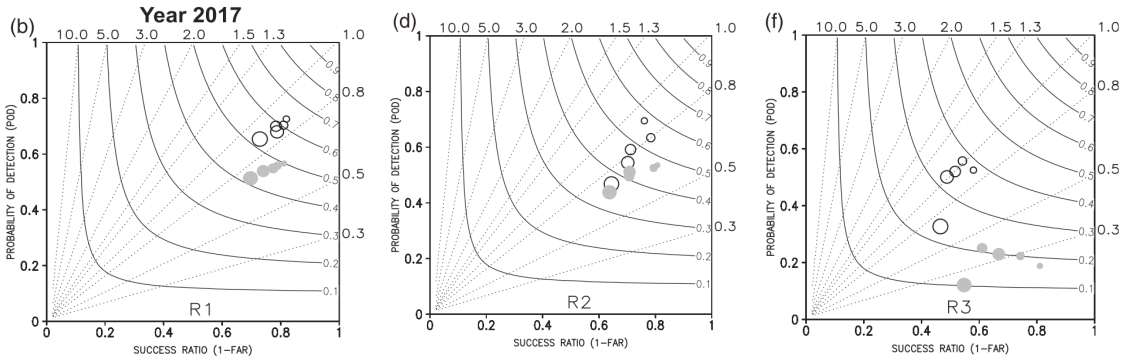
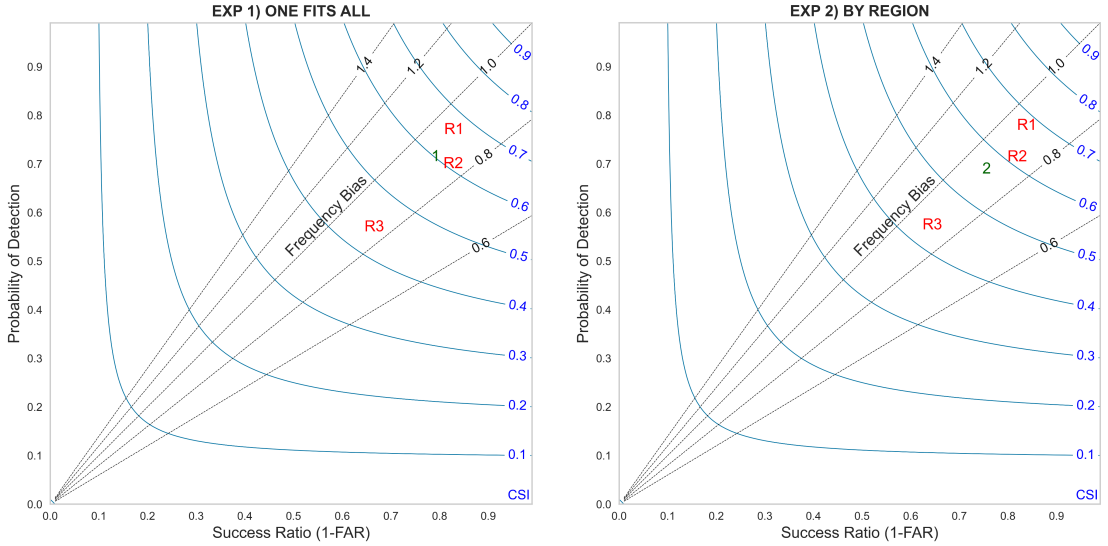


Figura 4. Resultado do desempenho do IG-CPTEC comparado ao modelo Eta, adaptado de Rozante et al. (2020, Figure 4)

A Figura 5 mostra um Diagramas de Desempenho para cada experimento deste trabalho: ONE-FIT-ALL (EXP 1), BY-REGION (EXP 2), CLUST-CLIM (EXP 3), e CLUST-LOC (EXP 4). Para todos os gráficos: O número em verde escuro representa o desempenho global referente ao experimento (1 a 4); em vermelho é mostrado o desempenho das regiões do IG-CPTEC (R1, R2 e R3); e em verde claro o desempenho dos clusteres formados no experimento (quando pertinente).



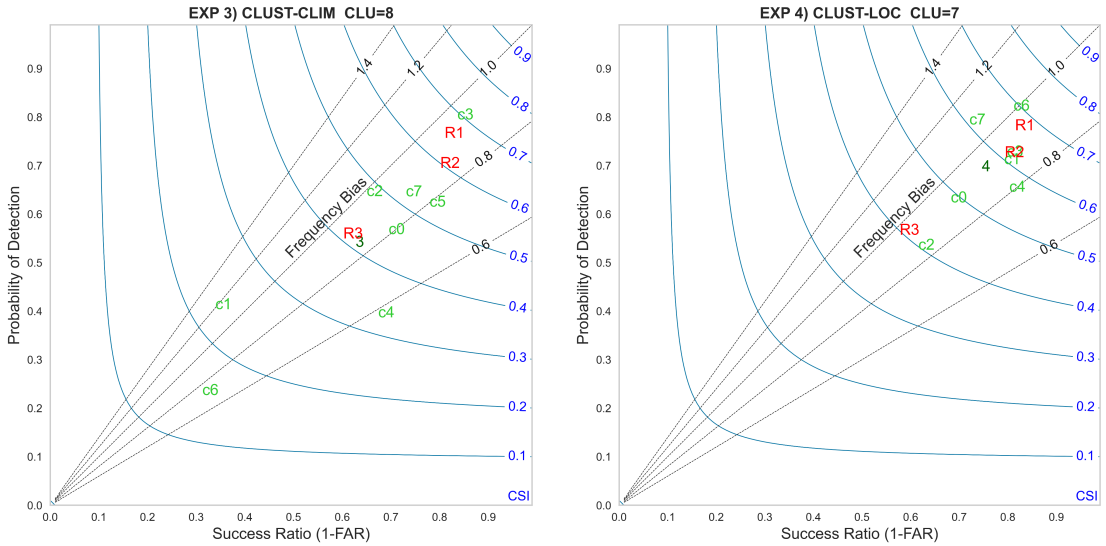


Figura 5. Resultado do desempenho dos experimentos 1 a 4

4. Conclusões

Usar aprendizado de máquina para a tarefa de previsão de geadas se mostra muito promissor. Os diagramas de desempenho mostraram que os quatro experimentos superam o IG-CPTEC para as previsões de 24h. Em relação ao desempenho global (números 1 a 4 em verde escuro), verifica-se que o ONE-FITS-ALL teve melhor desempenho, em seguida aparecem o BY-REGION e o CLUST-LOC praticamente idênticos. Por região, BY-REGION e CLUST-LOC desempenharam melhor que os demais nas regiões 1 e 2, enquanto que para a região 3 os experimentos BY-REGION e ONE-FITS-ALL se destacaram.

A região 3 apresenta-se sempre com desempenho inferior às demais, indicando que merece maiores estudos. O cálculo de importância dos preditores confirma que a temperatura é o aspecto que mais influencia na previsão de geadas, mas todos contribuem para a modelagem e não devem ser ignorados, conforme teste exploratório. Embora menos acentuado para a Região 3, os experimentos continuam tendendo a prever menos ocorrência de geadas como é o IG-CPTEC, que talvez seja explicado pelo fato de que desbalanceamento entre as ocorrências da classe 0 (sem geada) e da classe 1 (com geada) possa afetar a modelagem.

Os resultados motivam estudos futuros e mostram que não se pode descartar nenhuma das abordagens. Tarefas futuras incluem: extensão da metodologia para outros períodos de previsão; considerar outras variáveis preditoras, testar em outros *datasets* mais atuais e de outros modelos, estimar geada em multiclasses, usar modelos com capacidade de memória, fazer análise outras análises de relevância das variáveis, etc.

5. References

Aguiar, D., Mendonça, M. (2004) Climatologia das geadas em Santa Catarina. In: Simpósio Brasileiro de Desastres Naturais. Anais. Florianópolis.

Black, T.L. (1994) The new NMC mesoscale Eta model: description and forecast examples. Weather and Forecasting, 9, 265–278.

Chou, S.C., Tanajura, C.A.S., Xue, Y. and Nobre, C.A. (2002) Valida- tion of the coupled Eta/SsiB model over South America. Journal of Geophysical Research: Atmospheres, 107(D20), 8088. https://doi. org/10.1029/2000JD000270.

Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451

Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, 842 pp., Prentice-Hall, Old Tappan, N. J.

Liaw, A., Wiener, M., (2002) Classification and Regression by random forest. R. News 2, 18–22.

Mesinger, F., Janji?c, Z.I., Ni ckovi?c, S., Gavrilov, D. and Deaven, D.G. (1988) The step-mountain coordinate: model description and per- formance for cases of alpine lee cyclogenesis and for a case of an Appalachian redevelopment. Monthly Weather Review, 116, 1493–1518.

Naing, W.Y.N., Htike, Z.Z., (2015) Forecasting of monthly temperature variations using random forests. ARPN J. Eng. Appl. Sci. 10, 10109–10112.

Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

Pereira, A. R., Angelocci, L. R., Sentelhas, P. C. (2002) Agrometeorologia: fundamentos e aplicações práticas. Guaíba: Agropecuária. 478p.

Robinson, C., Mort, N., (1997) A neural network system for the protection of citrus crops from frost damage. Comput. Electron. Agric. 16, 177–187. https://doi.org/10.1016/ S0168-1699(96)00037-3.

Roebber, P.J. (2009) Visualizing multiple measures of forecast quality. Weather and Forecasting, 24, 601–608.

Rozante J.R., Gutierrez E.R, da Silva Dias P.L., de Almeida Fernandes A., Alvim D.S., Silva V.M. (2020) Development of an index for frost prediction: Technique and validation. Meteorol Appl. 2020;27:e1807. https://doi.org/10.1002/ met.1807

Stephenson, D.B. (2000) Use of the “odds ratio” for diagnosing fore- cast skill. Weather and Forecasting, 15, 221–232.

Wilks, D.S. (2011) Statistical methods in the atmospheric sciences. Oxford and Waltham, MA: Academic Press and Elsevier Science.

Yang L., Shami, A. (2020) On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing, Volume 415, Pages 295-316, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.07.061.