Comparação de algoritmos de aprendizado de máquina para predição de clima espacial: erupções solares e tempestades geomagnéticas

Fernando de S. Mello[1]

[1]Instituto Hardware BR

fdesmello@gmail.com

Abstract. *A set of algorithms (random forest, support vector machine, neural network, and recurrent neural network – GRU) was applied for space weather prediction: solar flares and geomagnetic storms. The positive prediction of as many events as possible was prioritized, even at the cost of more false positives, as these events are uncommon and potentially damaging. Random forest showed to be the worst algorithm, while the other three performed more closely, with neural networks generally standing out, but requiring more processing time. A few suggestions are given to improve the models: more and better features, multi-class prediction or regression, and the use of real-time data.*

Resumo. *Um conjunto de algoritmos (floresta aleatória, máquina de vetor de suporte, rede neural e rede neural recorrente – GRU) foi aplicado para predição de clima espacial: erupções solares e tempestades geomagnéticas. Foi priorizado a predição positiva de eventos verdadeiros, mesmo ao custo de mais falsos positivos, por serem eventos incomuns e potencialmente danosos. Floresta aleatória se mostrou o pior algoritmo, enquanto os outros três tiveram desempenho mais próximo e as redes neurais em geral se destacando, mas exigindo maior tempo de processamento. Algumas sugestões são dadas para melhorar os modelos: mais e melhores características, previsão multi-classe ou regressão e uso de dados em tempo real.*

1. Introdução

O clima espacial (também tempo espacial, *space weather*) refere-se às variações no ambiente espacial no Sistema Solar, mas com foco especial na interação entre o Sol e a Terra. (Cade III e Chan-Park, 2015; Denardini et al., 2016) O principal motor e agente do clima espacial é o Sol, principalmente por eventos de erupções solares, ejeções de massa coronal, vento solar, e tempestades geomagnéticas. A frequência desses eventos não é constante. O Sol tem estações de atividade seguindo um ciclo de aproximadamente 11 anos. Durante esse ciclo, é possível observar um lento aumento na quantidade de manchas solares, anomalias magnéticas e na frequência de erupções solares e ejeções de massa coronal; para então decair lentamente de novo.

Regiões ativas (*active regions*) são características temporárias na atmosfera solar que apresentam um forte e complexo campo magnético. Essas regiões são frequentemente observadas como manchas escuras (manchas solares ou *sunspots*), regiões mais frias na superfície solar que aparentam ser mais escuras do que a região circundante, durando de dias a meses. Essas regiões estão frequentemente associadas com grandes e súbitas

emissões de radiação eletromagnética (erupções solares ou *solar flares*), especialmente de radiação ultravioleta e X, durando da ordem de alguns minutos. Associadas às erupções solares há frequentemente a ejeção de grandes quantidades de plasma da corona solar (ejeção de massa coronal ou *coronal mass ejection* – CME). Esse plasma, quando cruza a órbita terrestre e interage com a magnetosfera de nosso planeta, provoca as auroras.

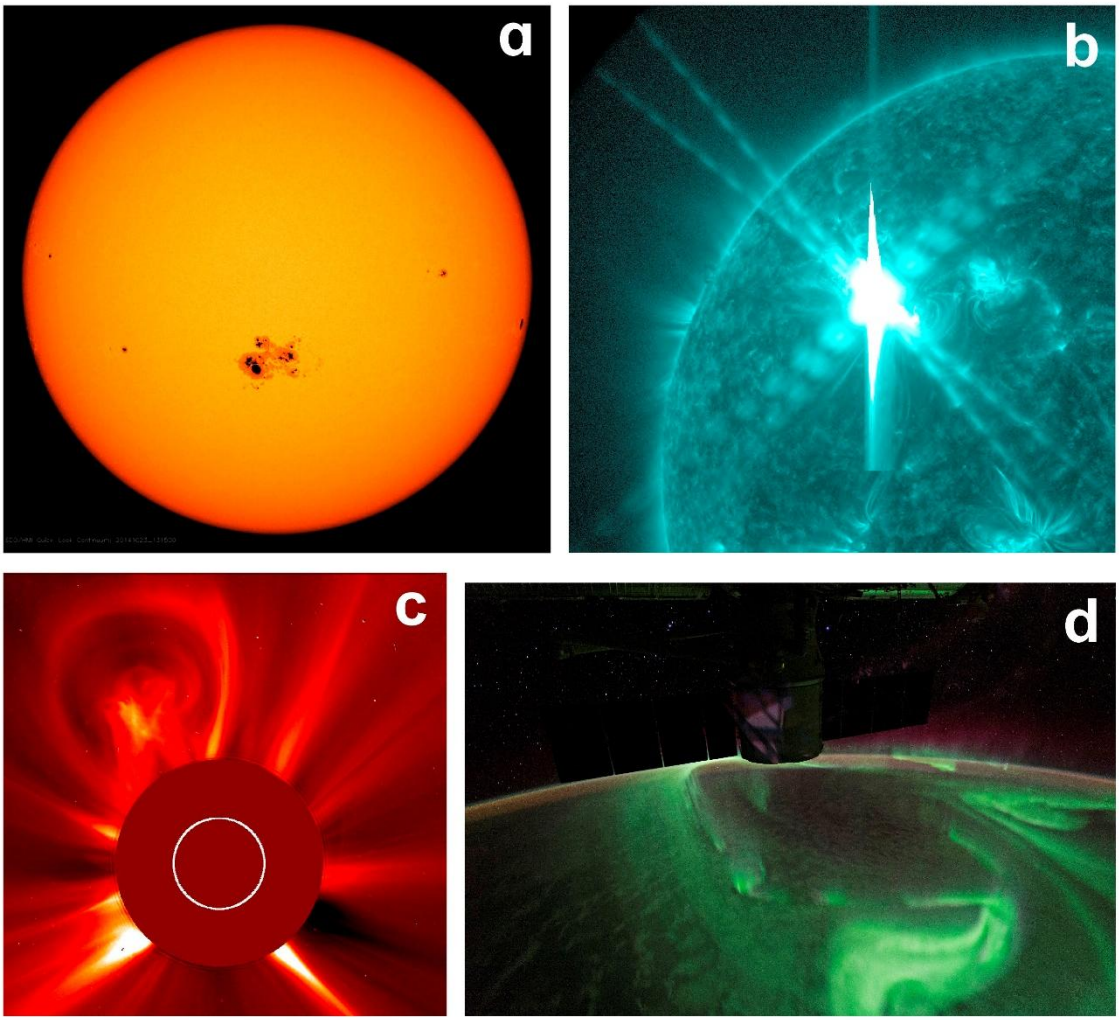

Figura 1. Eventos solares e terrestre relacionados com o clima espacial. a) Um grande conjunto de manchas solares. Crédito: NASA/SDO/AIA. b) Uma grande erupção solar vista em raios X. Crédito: NASA/SDO. c) Uma grande ejeção de massa coronal com coronógrafo ocultando a vista do Sol. Crédito: NASA/GSFC/SOHO. d) Aurora austral vista da Estação Espacial Internacional. Crédito: NASA.

Mas as auroras não são os únicos efeitos do clima espacial na Terra. A grande quantidade de radiação ionizante emitida durante uma erupção solar aumenta a ionização e temperatura da ionosfera terrestre, o que pode interferir com comunicação por rádio, sistemas de localização global, e aumentar o arrasto sentido por satélites na órbita terrestre baixa, pois a atmosfera superior se expande ao ser aquecida.

Ejeções de massa coronal, por sua vez, podem perturbar o campo magnético terrestre e provocar tempestades geomagnéticas (*geomagnetic storms*). As variações magnéticas e o grande fluxo de partículas carregadas podem afetar a comunicação e funcionamento de

satélites, causar picos nos sistemas de rede elétrica – podendo levar até a apagões – além de expor astronautas em órbita a maiores doses de radiação. (Adebesin et al., 2013)

Um exemplo recente de um evento intenso foi a tempestade solar de 1989, que causou interferência na comunicação por rádio, apagões no Canadá e o avistamento de auroras muito mais ao Sul do que de costume. (Boteler, 2019) No entanto, a maior tempestade solar já registrada foi a tempestade solar de 1859, também conhecida como Evento Carrington. (Tsurutani et al., 2003) Seu impacto prático se restringiu mais aos sistemas de telégrafo da época. Mas se uma tempestade semelhante acontecesse hoje, com muitos mais sistemas elétricos e eletrônicos do que no século XIX, o potencial destrutivo seria bem maior. (Phillips et al., 2014; Choi, 2024)

De maneira geral, esses eventos extremos do clima espacial podem ter consequências econômicas significativas e o potencial de afetar negativamente diversos setores da civilização humana. Daí a importância da previsão desses eventos para nos dar mais tempo de reação e mitigar seus efeitos potencialmente catastróficos.

Ao longo dos anos, vários modelos físicos foram usados para prever diferentes fenômenos do clima espacial. Mais recentemente, métodos de aprendizado de máquina (*machine learning*) também vêm sendo empregados com bons resultados devido ao aumento no poder de processamento de computadores e na grande quantidade de dados observacionais (Denardini et al., 2016). Até competições para desenvolver os melhores algoritmos vêm sendo realizadas, onde o prêmio, além de dinheiro, é ter o algoritmo utilizado para previsões reais. (Nair et al., 2023)

2. Metodologia

2.1. Objetivo

O objetivo do trabalho é criar algoritmos de predição para predizer o evento de interesse (erupção solar ou tempestade geomagnética) na próxima unidade de tempo (dia ou hora) com base nos dados das três últimas unidades de tempo (três dias ou três horas) observadas as limitações de tempo, processamento e experiência inerentes desse projeto.

Classificação foi decidido por ser um método comum na literatura para esse tipo de objetivo e é mais facilmente implementado com as limitações do projeto. O intervalo de tempo veio de testes iniciais, onde os dados mais úteis eram os imediatamente anteriores ao evento e usar uma janela de tempo longa não melhorava muito os resultados.

Na literatura, há trabalhos que tentam fazer previsões usando algoritmos de diferentes complexidades e obtêm resultados satisfatórios. Por isso, foi decidido realizar um panorama de quatro algoritmos abordados no curso: floresta aleatória (*random forest*, RF), máquina de vetor de suporte (*support vector machine*, SVM), redes neurais (*neural nets*, NN) e redes neurais recorrentes (*recurrent neural netwroks*, RNN) do tipo GRU (g*ated recurrent unit*). Os dois primeiros foram implementados usando a biblioteca Scikit-learn em Python, e os dois últimos, a biblioteca Tensorflow/Keras, também em Python.

Os dois primeiros algoritmos são comuns na literatura de previsão de erupções solares e os dois últimos na literatura de tempestades geomagnéticas. (Ribeiro e Gradvohl, 2021; Nair, 2023) Mas os quatro também oferecem a oportunidade de aplicar muito do aprendido no curso de uma vez e comparar os resultados dos quatro para um mesmo

problema e usando os mesmos dados. Não necessariamente o algoritmo mais complexo e demorado traria os melhores resultados.

2.2. Problemas e soluções

Erupções solares e tempestades geomagnéticas são fenômenos infrequentes, sazonais e de curta duração. Isso cria datasets potencialmente bastante desbalanceados. A natureza temporal dos dados exige cuidado especial tanto em sua ordenação quanto em evitar o vazamento de dados (*data leakage*). São fenômenos com grande potencial de danos, e os eventos mais energéticos – de maior interesse – são também os mais raros, isso torna importante a previsão do maior número possível de eventos mesmo que ao custo de alguns falsos positivos.

A temporalidade foi tratada deslocando o alvo (*target*) uma linha para trás e separando o dataset em treino, validação e teste em fatias contínuas cronologicamente em vez de aleatoriamente, o que misturaria os dados dentro da janela de tempo de interesse. Como os eventos acontecem sazonalmente ao longo do ciclo solar de 11 anos, foi preciso usar um longo intervalo de tempo e selecionar bem o intervalo de dados dos datasets para que não houvesse falta de eventos em qualquer parte da separação treino, validação e teste.

A janela de tempo de informação foi tratada de maneira diferente dependendo do algoritmo. Para RF, SVM e NN, triplicou-se o número de características ao criar cópias atrasadas das características na janela de tempo de interesse. Isso transforma uma sequência de linhas em uma sequência de colunas ou características (*features*). Isso ajuda na questão de temporalidade, pois toda a informação necessária para predição está na mesma linha (caso), incluindo o alvo. Para RNN-GRU foi usada uma janela deslizante de tempo e não foi preciso multiplicar as características, e a ordenação temporal das linhas se manteve ainda mais importante.

O desbalanceamento foi inicialmente tratado com a criação de dados sintéticos baseados na classe minoritária usando SMOTE (*Synthetic Minority Over-sampling Technique*). No entanto, devido ao maior tempo de processamento e o temor da possibilidade de distorção da série histórica, foi escolhido o método de pesos, dando maior peso de maneira inversa à frequência da classe nos datasets.

Também dado o desbalanceamento nos dados, a necessidade de prever a classe minoritário, e que perder um evento raro poderia ter consequências danosas, as métricas utilizadas teriam de focar em diminuir falsos negativos, nesse caso, revocação e F_2 score (F_2). Outras métricas foram calculadas também para fim de comparação: F_1 score (F_1), precisão média (*average precision*), e área sob a curva característica de operação do receptor (*area under the receiver operating characteristic curve*, ROC-AUC).

Por fim, a escolha de hiperparâmetros dos algoritmos foi feita na etapa de validação usando busca em grid para RF e SVM, e busca aleatória para NN e RNN-GRU. Veja os hiperparâmetros finais na Tabela 4.

Os modelos NN foram constituídos de 4 camadas:

1. Dense (função de ativação: ReLU)
2. Dropout
3. Dense (função de ativação: ReLU)
4. Dense (função de ativação: Sigmoide)

Os modelos RNN-GRU foram constituídos de 5 camadas:

1. GRU (função de ativação: Tangente hiperbólica)
2. Dropout
3. Dense (função de ativação: ReLU)
4. Dropout
5. Dense (função de ativação: Sigmoide)

2.3. Dados primários

Os dados primários foram obtidos de três principais fontes:

- PyPEDAS, uma biblioteca Python que oferece acesso a muitos dados astronômicos. (The SunPy Community et al., 2020) Em especial, foram usados dados dos projetos:
    - GOES, uma série de satélites especialmente projetados para colher dados solares.
    - OMNI, um conjunto de dados relativos ao Sol oriundos de diferentes fontes.
- DRMS, outra biblioteca Python para acesso a uma grande variedade de dados relacionados ao clima espacial. (Glogowski et al., 2019) Em especial, foram usados dados do SHARP, uma série de quantidades calculadas de dados fotométricos solares. (Bobra et al., 2014)
- Kaggle, plataforma e comunidade online de competição de ciência de dados que frequentemente disponibiliza modelos e datasets. (erevear, 2024)

Os dados obtidos do OMNI contêm basicamente informações a respeito de campo magnético e plasma no meio sideral obtido de satélites no ponto Lagrangiano L_1 entre a Terra e o Sol e foram usados no dataset para tempestades geomagnéticas. Os outros datasets contêm uma variedade maior de dados, mas se concentram em fluxo de radiação e campo magnético de regiões ativas do Sol obtidos de satélites. Esses foram mesclados no dataset para as erupções solares.

2.4. Limpeza dos datasets

O datasets finais passaram por processos semelhantes de limpeza e redução que consumiu muito do tempo do projeto.

A resolução temporal dos datasets iniciais para erupções solares não era a mesma e foi preciso tirar a média de valores ao longo de 24 horas para mesclá-los adequadamente. Nesses casos, a média foi feita também sobre várias regiões ativas do Sol naquele momento e ao longo de um dia. A perda de resolução e informação pode ter sido importante e isso será comentado mais à frente. Além da média, foi também incluído os valores máximo e mínimo. Algumas colunas de dados (potenciais características) continham até mais de 70% de valores nulos e foram excluídas. Buracos (valores nulos) nas colunas de até 3 dias consecutivos foram interpolados e buracos de até 6 dias foram preenchidos com a média móvel. Buracos maiores levaram a exclusão das linhas. Isso cria falta de continuidade em alguns locais do dataset, mas não em tantos locais a ponto de ser relevante.

O dataset para tempestades geomagnéticas passou por procedimento análogo, só não foi preciso a normalização de resolução temporal.

A escolha das características de ambos os datasets foi feita de maneira conservativa. Já se sabia de antemão algumas informações que seriam importantes incluir, como campo magnético, manchas solares e fluxo de raios X. Mas, além disso, as características escolhidas foram as mais usadas na literatura para esse tipo de trabalho. Como as características abarcam muitas ordens de grandeza, foi aplicado uma normalização nas características, mas apenas logo antes dos ajustes, não nos datasets finais em si.

2.5. Os alvos

Erupções solares são classificadas segundo o máximo de emissão de raios X no comprimento de onda de 0,1 a 0,8 nm. (Messerotti et al., 2009)

Tabela 1. Classificação de erupções solares.

Classe	Pico de fluxo (W m^{-2})
A	$< 10^{-7}$
B	$10^{-7} – 10^{-6}$
C	$10^{-6} – 10^{-5}$
M	$10^{-5} – 10^{-4}$
X	$> 10^{-4}$

As erupções mais fortes, M e X, foram consideradas de maior interesse por influenciarem de maneira mais sensível as atividades terrestres e foi criada uma nova característica binária no dataset para assinalar isso: flare_today, indicando um dia com ao menos um evento de qualquer uma das duas classes de erupções. Ao redor de 12% dos dias no dataset contêm esse tipo de evento. Esse é o alvo de predição para os classificadores binários.

Tempestades geomagnéticas são classificadas segundo o índice de distúrbio-tempo de tempestade (*disturbance storm-time index*, Dst), um índice de atividade magnética derivado de uma rede de observatórios geomagnéticos na Terra. Quanto maior a diferença entre o observado e o valor típico, mais fraco o campo magnético e maior a interferência magnética pela tempestade geomagnética. Valores típicos variam de 20 a -20 nT. Durante uma tempestade, o índice assume valores inferiores a -50 nT. (Adebesin et al. 2013) Esse limite é algo arbitrário, pois o índice é uma variável contínua e não há um sentido físico especial para esse ponto.

Tabela 2. Classificação de tempestades geomagnéticas em função da perturbação no índice Dst.

Classe de tempestade	Valor mínimo do índice Dst (nT)
Fraca ou Ausente	> -50
Moderada	-50 – -100
Forte	-100 – -200
Severa	-200 – -350
Grande	< -350

Índices com valores Dst abaixo de -50 nT foram considerados como sinal de uma tempestade e foi criada uma nova característica binária no dataset para assinalar isso: storm_now, indicando uma tempestade naquela hora. Ao redor de 3.5% das horas no

dataset contêm esse tipo de evento. Esse é o alvo de predição para os classificadores binários.

Em ambos os casos, há o risco de misturar diferentes tipos de eventos com tal classificação simples, mas como eventos de classes mais altas são mais raros, isso pode não importar tanto assim inicialmente. Mais sobre isso será discutido adiante.

2.6. Os datasets finais

2.6.1. Erupções solares

O dataset final para erupções solares contêm 33 colunas e 4480 linhas, com resolução na escala de dia, indo de maio de 2010 a janeiro de 2023.

Tabela 3. Informações do dataset para erupções solares. Fluxos e campos quando não especificados são magnéticos e corrente é corrente elétrica. *: Excluído de uso como característica.

Característica	Unidade	Descrição
DATE		Data
USFLUX	Mx	Total de fluxo magnético
MEANGAM	°	Ângulo médio de campo em relação ao radial
MEANGBT	G Mm^{-1}	Gradiente horizontal do campo total
MEANGBZ	G Mm^{-1}	Gradiente horizontal do campo vertical
MEANGBH	G Mm^{-1}	Gradiente horizontal de campo horizontal
MEANJZD	mA m^{-1}	Densidade de corrente vertical
TOTUSJZ	A	Total de corrente vertical
MEANALP	Mm^{-1}	Parâmetro de torção característico, α
MEANJZH	G^2 m^{-1}	Helicidade da corrente (contribuição de Bz)
TOTUSJH	G^2 m^{-1}	Helicidade total da corrente
ABSNJZH	G^2 m^{-1}	Valor absoluto da helicidade de corrente líquida
SAVNCPP	A	Soma do módulo da corrente líquida por polaridade
MEANPOT	erg cm^{-3}	Indicador da densidade média de excesso de energia magnética na fotosfera
TOTPOT	erg cm^{-3}	Indicador da densidade total de energia livre magnética da fotosfera
MEANSHR	°	Ângulo de cisalhamento
SHRGT45		Área fracionária com cisalhamento >45°
R_VALUE	Mx	Fluxo magnético próximo às linhas de inversão de polaridade
xrs_A_mean	W m^{-2}	Média de fluxo de raios X de curto comprimento de onda
xrs_A_min	W m^{-2}	Mínimo de fluxo de raios X de curto comprimento de onda
xrs_A_max	W m^{-2}	Máximo de fluxo de raios X de curto comprimento de onda
xrs_B_mean	W m^{-2}	Média de fluxo de raios X de longo comprimento de onda
xrs_B_min	W m^{-2}	Mínimo de fluxo de raios X de longo comprimento de onda
xrs_B_max	W m^{-2}	Máximo de fluxo de raios X de longo comprimento de onda
Radio Flux 10.7cm	s.f.u. (10^{-22} W m^{-2} Hz^{-1})	Fluxo de rádio em comprimento de onda de 10,7 cm
Sunspot Number		Número de manchas solares

Sunspot Area	MSH (10^{-6} hemisfério)	Área somada das manchas solares
New Regions		Número de novas manchas solares
*Flares: C		Número de erupções classe C
*Flares: M		Número de erupções classe M
*Flares: X		Número de erupções classe X
flare_today		*Flag* para erupções
flare_missing		*Flag* para algum dado faltando na fonte

2.6.2. Tempestades geomagnéticas

O dataset final para tempestades geomagnéticas contêm 19 colunas e 244.535 linhas, com resolução na escala de hora, indo de janeiro de 1995 a dezembro de 2024.

Tabela 4. Informações do dataset para tempestades geomagnéticas. Componentes e campos são magnéticos. Quando não especificado, outras grandezas são referentes ao plasma. *: Excluído de uso como característica.

Característica	Unidade	Descrição
datetime		Data
ABS_B	nT	Magnitude do vetor de campo médio, |<B>|
F	nT	Magnitude média do campo, <F>
BX_GSE	nT	Componente X do campo magnético interplanetário usando o sistema de coordenadas GSE
BY_GSE	nT	Componente Y do campo magnético interplanetário usando o sistema de coordenadas GSE
BZ_GSE	nT	Componente Z do campo magnético interplanetário usando o sistema de coordenadas GSE
SIGMA-ABS_B	nT	Desvio Padrão RMS na magnitude média
SIGMA-B	nT	Desvio Padrão RMS no vetor de campo
SIGMA-Bx	nT	Desvio Padrão RMS da componente X média no GSE
SIGMA-By	nT	Desvio Padrão RMS da componente Y média no GSE
SIGMA-Bz	nT	Desvio Padrão RMS da componente Z média no GSE
T	K	Temperatura
N	cm^{-3}	Densidade
V	km s^{-1}	Velocidade
Ratio		Razão entre partículas alfa e prótons
Pressure	nPa	Pressão do fluxo
R		Número de novas manchas solares por dia
*DST	nT	Índice Dst
storm_now		*Flag* de hora com tempestade

3. Resultados

Para mais detalhes, por favor veja código e repositório no GitHub. Há mais figuras e dados que não couberam aqui.

Tabela 5. Hiperparâmetros para os quatro algoritmos utilizados e para os dois fenômenos previstos.

Erupções solares

RF	SVM	NN	RNN-GRU
max depth: 10	C: 1	layer 1 units: 128	units gru: 64
max features: None	gamma: 0,001	layer 2 units: 256	units dense: 32
min samples leaf: 10	kernel: rbf	dropout rate: 0,3	dropout rate: 0,3
min samples split: 2	scoring: F2	learning rate: 0,0001	learning rate: 0,0001
n estimators: 50		batch size: 128	batch size: 128
scoring: F2		loss: binary crossentropy	loss: binary crossentropy

Tempestades geomagnéticas

RF	SVM	NN	RNN-GRU
max depth: 10	C: 0,1	layer 1 units: 512	units gru: 32
max features: log2	gamma: 0,001	layer 2 units: 256	units dense: 32
min samples leaf: 10	kernel: rbf	dropout rate: 0,5	dropout rate: 0,5
min samples split: 2	scoring: F2	learning rate: 0.01	learning rate: 0,0001
n estimators: 200		batch size: 128	batch size: 128
scoring: F2		loss: binary crossentropy	loss: binary crossentropy

Tabela 6. Métricas calculadas para os quatro algoritmos utilizados e para os dois fenômenos previstos.

Erupções solares

	F_2	F_1	Revocação	Precisão média	ROC-AUC
RF	0,518	0,452	0,574	0,448	0,825
SVM	0,614	0,476	0,762	0,434	0,824
NN	0,555	0,435	0,680	0,422	0,820
RNN-GRU	0,636	0,472	0,828	0,537	0,854

Tempestades geomagnéticas

	F_2	F_1	Revocação	Precisão média	ROC-AUC
RF	0,449	0,411	0,478	0,410	0,917
SVM	0,429	0,254	0,792	0,355	0,895
NN	0,450	0,284	0,737	0,477	0,930
RNN-GRU	0,436	0,243	0,928	0,511	0,959

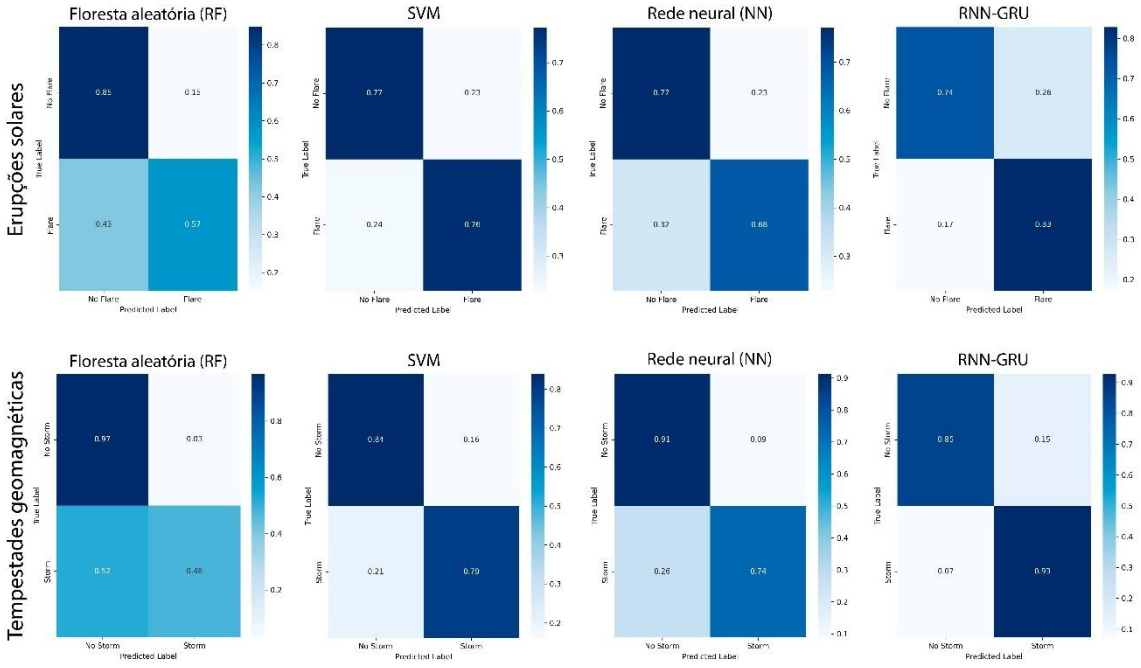

Figura 2. Matrizes de confusão normalizadas por classes verdadeiras.

4. Análise e discussão dos resultados

Nos algoritmos de floresta aleatória, as características mais importantes foram em média as de atraso 1, seguidas por atraso 2 e então 3. Isso quer dizer que os dados recentes, logo antes do momento predito, são mais importantes. Alongar a janela de tempo de informação por mais de 3 dias não melhoraria substancialmente a predição e as métricas. Para erupções solares, as características mais importantes foram as relacionadas às emissões em raios X, seguidas por medidas de campo magnético das regiões ativas solares. No caso de tempestades geomagnéticas, as características mais importantes foram relacionadas ao campo magnético médio e medidas gerais do plasma. Ambos fazem sentido físico, com as principais grandezas possivelmente aumento de valor antes dos eventos.

De maneira geral e considerando testes iniciais, uma maior quantidade de dados no dataset para treino levou a resultados melhores nas métricas, mas também exigiu muito mais tempo de processamento para seleção de hiperparâmetros. Isso foi parcialmente resolvido para os três primeiros algoritmos selecionando uma fração do dataset de tempestades geomagnéticas (muito maior do que de erupções solares) na etapa de ajuste de hiperparâmetros, mas usando o dataset inteiro quando ajustando o modelo final. Para o algoritmo RNN-GRU, que usa uma janela temporal deslizando, isso não foi possível. Ainda assim, os tempos médios de seleção de hiperparâmetros para RF e SVM passaram de ~5 min em erupções solares para ~13 min em tempestades geomagnéticas; NN foi de ~20 minutos para ~4 horas; e RNN-GRU, de ~30 minutos para ~8 horas.

Fazendo variações do número de eventos como alvo (variando as classes de erupções incluídas ou o limite Dst para tempestade) e classificações erradas aumentam se há poucos eventos, mesmo que mais fortes, e diminuem com mais eventos, mesmo que mais fracos. Isso também fez ser mais difícil predizer corretamente eventos longe do pico do ciclo solar, pois há menos eventos.

Quando se usa mais de uma métrica, obtém-se um melhor panorama do desempenho dos algoritmos, mas é difícil ranqueá-los de maneira única. Dessa forma, os resultados foram mistos, baixos valores F_2, mas alguns valores aceitáveis para revocação. Para erupções solares, o pior algoritmo em geral foi floresta aleatória, enquanto RNN-GRU foi o melhor. Para tempestades geomagnéticas, floresta aleatória se destaca mais, mas tem uma revocação pobre, e RNN-GRU parece ser o melhor. Em ambos, SVM surpreende com desempenho intermediário.

Comparando as matrizes de confusão, um ranque mais claro parece surgir. Para Ambos os casos, SVM e NN tiveram desempenho semelhante na classe minoritária e na majoritária. Floresta aleatória mostra os piores resultados; e RNN-GRU, os melhores. Pode ser que estruturas úteis para predição contidas nos dados sejam mais facilmente percebidas por algoritmos mais complexos do que mais simples. Mas é curioso que a diferença não seja tão grande entre os três algoritmos de melhor desempenho.

As colunas Flares: C/M/X e índice Dst não foram utilizadas de características. Isso se deu por receio de estarem atreladas demais aos eventos sendo preditos. Muitas erupções solares acontecem próximas no tempo, a existência de alguma erupção anterior poderia sinalizar uma próxima. E tempestades geomagnéticas duram por algumas horas então a existência de uma hora marcada como tempestade indicaria que as seguintes provavelmente também seriam. Isso não necessariamente seria vazar informação do futuro, seria uma inferência lógica. Mas o interesse maior na predição é em predizer os eventos antes deles ocorrem, não quando já estão em curso. Outro motivo mais importante seria que essas características poderiam não estar disponíveis em tempo real, então não serviriam para predição. (Nair et al., 2023) A inclusão dessas características no treinamento aumentou o desempenho dos algoritmos (Dst mais do que Flares: C/M/X), mas poderia não ser algo muito útil mais tarde, quando os modelos estiverem sendo usados para previsão real.

4.1. Próximos passos

Os resultados são mistos, mas alguns algoritmos se destacam, embora ainda distantes do estado da arte do meio. Alguns melhoramentos podem diminuir essa distância. Esses são separados em três categorias: características, alvos e tempo real.

Ao algo pode ter sido perdido ao tirar as médias sobre 24 horas e sobre as regiões ativas. Criar características que retenham algo da informação anterior – como crescimento e decrescimento, valores máximos e mínimos – poderia ser útil. Escolher as regiões ativas mais promissoras e de campo magnético mais complexo antes de tirar as médias poderia ajudar também. (Hazra et al., 2020; Li et al., 2025) Assim como incluir características excluídas inicialmente por não serem populares na literatura. (Deshmukh et al., 2023) Mais características não necessariamente vão melhorar o resultado, mas pode haver alguma informação útil a ser adicionada.

Os alvos de predição, sendo binários, podem ser muitos simples para abarcar a complexidade dos fenômenos envolvidos. Seria interessante passar de um algoritmo classificador binário para um classificador de multi-classes. Como erupções solares de diferentes classes podem acontecer no mesmo dia, seria preciso escolher entre predizer a classe mais energética ou todas as diferentes classes daquele dia. E como o índice Dst é uma variável contínua, poderia ser melhor usar o próprio índice Dst de alvo e trocar o algoritmo classificador binário por um algoritmo de regressão.

Os modelos foram treinados usando dados consolidados. Dados em tempo real são mais ruidosos e menos confiáveis do que os consolidados. Algumas quantidades podem nem mesmo estar disponíveis até bem depois que o evento acabou. Dessa forma, se o objetivo for usar os modelos para predizer os eventos em tempo real, seria preciso testar em condições mais realistas e com o fluxo de dados real. Isso pode exigir algumas modificações quanto às características utilizadas e robustez no pré-processamento dos dados. Mas seria algo importante para tornar a previsão mais útil como alerta contra eventos extremos.

5. Conclusões

Um conjunto de algoritmos (floresta aleatória, máquina de vetor de suporte, rede neural e rede neural recorrente do tipo GRU) foram utilizados como classificadores binários para a predição de erupções solares e tempestades geomagnéticas, visando futuramente servir de sistema de alerta contra esses eventos. Como esses são eventos infrequentes, sazonais, e potencialmente danosos, foi priorizado a predição de positivos verdadeiros, mesmo ao custo de mais falsos positivos. Na comparação dos resultados, floresta aleatória se mostrou o pior algoritmo, enquanto os outros três tiveram desempenho mais próximo e aceitável, com RNN-GRU em geral se destacando, mas exigindo maior tempo de processamento.

Os resultados evidenciam a exequibilidade do procedimento com poucos recursos e o potencial da previsão do clima espacial, em especial de erupções solares e tempestade geomagnéticas. Mas vários melhoramentos ao procedimento ainda são possíveis.

Referências

Adebesin, B.O., Ikubanni, S.O., Kayode, J.S., Adekoya, B.J. (2013) "Variability of Solar Wind Dynamic Pressure with Solar Wind Parameters During Intense and Severe Storms", Em: African Review of Physics (International Centre for Theoretical Physics, ICTP, Italy].), Volume 8.

Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., Barnes, G., Leka, K. D. (2014) "The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches", Em: Solar Cycle 24 as seen by SDO, Volume 289, páginas 3549–3578.

Boteler, D. H. (2019) "A 21st Century View of the March 1989 Magnetic Storm", Em: Space Weather, Volume 17, Edição 10, pp. 1427-1441.

Cade III, W. B., Chan-Park, C. (2015) "The Origin of “Space Weather”", Em: Space Weather, Volume 13, Edição 2, pp. 99-103.

Choi, C. Q. (2024) "What if the Carrington Event, the largest solar storm ever recorded, happened today?", https://www.livescience.com/carrington-event. Acesso em: 2 de setembro de 2025.

Denardini, C. M., Dasso, S., Gonzalez-Esparza, J. A. (2016), Em: Review on space weather in Latin America. 2. The research networks ready for space weather. Volume 58, Edição 10.

Deshmukh, V., Baskar, S., Berger, T. E., Bradley, E., Meiss, J. D. (2023) "Comparing feature sets and machine-learning models for prediction of solar flares: Topology,

physics, and model complexity", Em: Astronomy and Astrophysics, Volume 674, A159.

erevear (2024) “Space Weather: Solar + Geomagnetic Indices” [Dataset]. Kaggle. Disponível em: https://www.kaggle.com/datasets/erevear/space-weather-solar-geomagnetic-indices. Acesso em: 26 de setembro de 2025.

Glogowski, K, Bobra, M. G., Choudhary, N., Amezcua, A. B., Mumford, S. J. (2019) "drms: A Python package for accessing HMI and AIA data", Em: Journal of Open Source Software, 4(40), 1614.

Hazra, S., Sardar, G., Chowdhury, P. (2020) "Distinguishing between flaring and nonflaring active regions", Em: Astronomy and Astrophysics, Volume 639, A44.

Li, X., Li, X., Zheng, Y., Li, T., Yan, P., Ye, H., Zhang, S., Wang, X., Lv, Y., Huang, X. (2025) "Prediction of Large Solar Flares Based on SHARP and High-energy-density Magnetic Field Parameters", Em: The Astrophysical Journal Supplement Series, 276:7 (14pp).

Phillips, T. (2014) "Near Miss: The Solar Superstorm of July 2012", https://science.nasa.gov/science-research/planetary-science/23jul_superstorm/. Acesso em: 2 de setembro de 2025.

Ribeiro, F. e Gradvohl, A.L.S. (2021) "Machine learning techniques applied to solar flares forecasting". Em: Astronomy and Computing, Volume 35.

Messerotti, M., Zuccarello, F., Guglielmino, S. L., Bothmer, V., Lilensten, J., Noci, G., Storini, M., Lundstedt, H. (2009) "Solar Weather Event Modelling and Prediction", Em: Space Science Reviews, Volume 147, páginas 121–185.

Nair, M., Redmon, R., Young, L., Chulliat, A., Trotta, B., Chung, C., Lipstein, G., Slavitt, I. (2023) "MagNet—A Data-Science Competition to Predict Disturbance Storm-Time Index (Dst) From Solar Wind Data", Em: Space Weather, Volume 21, Edição 10.

The SunPy Community et al. (2020) "The SunPy Project: Open Source Development and Status of the Version 1.0 Core Package", Em: The Astrophysical Journal, Volume 890, Número 1.

Tsurutani, B. T., Gonzalez, W. D., Lakhina, G. S., Alex, S. (2003) "The extreme magnetic storm of 1–2 September 1859", Em: Journal of Geophysical Research: Space Physics, Volume 108, Edição A7.