UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Renan de Oliveira Ferretti

Predição do Tempo de Trânsito de uma CME até a Terra

Limeira
2023

Renan de Oliveira Ferretti

Predição do Tempo de Trânsito de uma CME até a Terra

Dissertação apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Bacharel em Sistemas de Informação, na área de Sistemas de Informação e Comunicação.

Orientador: Prof. Dr. André Leon Sampaio Gradvohl

Este trabalho corresponde à versão final da Dissertação defendida por Renan de Oliveira Ferretti e orientada pelo Prof. Dr. André Leon Sampaio Gradvohl.

Limeira
2023

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Bacharel em Sistemas de Informação na área de concentração Sistemas de Informação e Comunicação, a que se submeteu o aluno Renan de Oliveira Ferretti, em 16 de junho de 2023 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Prof. Dr. André Leon Gradvohl
Presidente da Comissão Julgadora

Prof. Dr. João Roberto Bertini Júnior
FT/UNICAMP

Prof. Dr. Celmar Guimarães da Silva
FT/UNICAMP

Ata da defesa, assinada pelos membros da Comissão Examinadora, encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação da Faculdade de Tecnologia.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a realização deste trabalho. Sem o apoio e incentivo de pessoas especiais, essa conquista não seria possível. Portanto, gostaria de dedicar esta seção de agradecimentos para expressar minha sincera gratidão a cada um de vocês.

À minha família, meus pais, Soraya e Wladimir, minha cachorrinha, Shakira, meus avós, tios e primos, que sempre acreditaram em mim, apoiaram meus sonhos e me incentivaram a buscar a excelência em meus estudos, não há palavras suficientes para expressar minha gratidão. Vocês estiveram ao meu lado em todos os altos e baixos, me fornecendo um ambiente amoroso e encorajador que me deu confiança para enfrentar qualquer desafio. Agradeço por todo o amor, paciência e apoio incondicional ao longo desta jornada. Alguns não estão mais aqui e espero estar honrando a memória de vocês, conforme prometido. Amo todos vocês.

Aos meus amigos, que estiveram ao meu lado durante toda essa jornada acadêmica, compartilhando risadas, frustrações e momentos inesquecíveis, eu sou imensamente grato. Vocês foram meu suporte emocional e minha força motivadora em momentos desafiadores.

Aos meus professores, que compartilharam seu conhecimento e experiência, e dedicaram seu tempo para me orientar e inspirar, sou profundamente grato. Suas aulas foram mais do que meras lições; foram verdadeiras fontes de inspiração. Agradeço por suas palavras sábias, por desafiarem meus limites intelectuais e por me encorajarem a explorar meu potencial ao máximo. Sou grato pela maneira como vocês moldaram meu pensamento crítico e minha paixão pelo aprendizado.

Ao meu orientador, cuja orientação, apoio e paciência foram fundamentais para a conclusão deste trabalho, expresso minha mais sincera gratidão. Sua expertise e visão crítica foram inestimáveis para o desenvolvimento deste projeto. Obrigado por compartilhar seu conhecimento, por sua dedicação incansável e por me inspirar a dar o melhor de mim. Sou imensamente grato pela confiança que você depositou em mim e pela oportunidade de aprender com você.

Por fim, gostaria de agradecer à minha universidade, que me proporcionou um ambiente estimulante e recursos necessários para minha formação acadêmica. Oportunidades de aprendizado, infraestrutura de qualidade e uma comunidade acadêmica vibrante foram fatores essenciais para meu crescimento pessoal e profissional. Sou grato por ter a oportunidade de fazer parte dessa instituição de ensino e pela formação que recebi ao longo dos anos.

A todos que mencionei e àqueles que me apoiaram de outras maneiras, mesmo que não mencionados aqui, meu sincero agradecimento. Cada palavra de encorajamento, gesto amável e suporte emocional contribuíram para minha jornada e tornaram este momento possível. A todos vocês, minha eterna gratidão.

Resumo

O clima espacial é o estudo das condições e fenômenos do espaço que podem afetar as atividades humanas e tecnológicas na Terra, incluindo tempestades solares, ejeções de massa coronal, variações no campo magnético e radiação cósmica. Nesse contexto, buscamos evitar ou mitigar os efeitos que esses fenômenos causam em nossas tecnologias. Para isso existem diversos modelos que tentam realizar predições do tipo de fenômeno, sua gravidade e quando vamos sentir os seus efeitos em nosso planeta. Desse modo, a previsão do tempo de trânsito de um fenômeno desempenha um papel crucial na mitigação dos impactos potenciais desses eventos na tecnologia espacial e nas comunicações terrestres. Neste trabalho de conclusão de curso, propomos um estudo comparativo de três modelos de predição do tempo de trânsito da ejeção de massa coronal com base em dados observacionais e em técnicas de aprendizado de máquina. Inicialmente, coletamos um conjunto de dados de ejeções de massa coronal previamente registradas por observatórios. Esses dados incluem parâmetros físicos relevantes, como velocidade inicial, massa e orientação da ejeção de massa coronal. Em seguida, utilizamos algoritmos de aprendizado de máquina para treinar e validar nosso modelo de predição. Mais especificamente, utilizamos uma rede neural do tipo *Multi-Layer Perceptron*, uma rede neural do tipo *Long Short Term Memory*, e um conjunto de preditores contendo os modelos *Support Vector Regression*, *Gaussian Process Regressor* e o XGB *Regressor*. Utilizamos as principais métricas que vêm sendo empregadas no estudo do tempo de trânsito de uma CME para avaliar cada modelo, como o erro médio absoluto. Os resultados obtidos mostraram que esse conjunto de preditores foi superior aos outros dois modelos, se aproximando do desempenho de dois dos principais modelos da literatura.

Palavras-chave: aprendizado de máquina, predição, ejeção de massa coronal, séries temporais, clima espacial

Abstract

Space weather is the study of space conditions and phenomena that can affect human and technological activities on Earth, including solar storms, coronal mass ejections, magnetic field variations, and cosmic radiation. In this context, we seek to avoid or mitigate the effects that these phenomena have on our technologies. For this, there are several models that try to make predictions about the type of phenomenon, its severity and when we will feel its effects on our planet. Thus, the prediction of the transit time of a phenomenon plays a crucial role in mitigating the potential impacts of these events on space technology and on terrestrial communications. In this course completion work, we propose a comparative study of three models for predicting the transit time of coronal mass ejection based on observational data and machine learning techniques. Initially, we collected a dataset of coronal mass ejections previously recorded by observatories. These data include relevant physical parameters such as initial velocity, mass and coronal mass ejection orientation. We then use machine learning algorithms to train and validate our prediction model. More specifically, we use a neural network of the type *Multi-Layer Perceptron*, a neural network of the type *Long Short-Term Memory*, and a set of predictors containing the models *Support Vector Regression*, *Gaussian Process Regressor* and the XGB *Regressor*. We used the main metrics that have been used in the study of the transit time of a CME to evaluate each model, such as the mean absolute error. The results obtained showed that this set of predictors was superior to the other two models, approaching the performance of two of the main models in the literature.

Keywords: machine learning, prediction, coronal mass ejection, time series, space weather

Lista de Figuras

2.1 Média anual de manchas solares do tipo Wolf e de grupo entre os anos 1610-2000. Disponível em (EDDY, 1976). 16
18figure.2.2
2.3 Funcionamento do Perceptron. Disponível em Ramchoun et al. (2016). 23
2.4 Estado inicial da janela deslizante. 26
2.5 Estado da janela após deslizar. 26

3.1 Seleção de atributos que servirão de entrada para os modelos. 33
3.2 Construção da MLP através da biblioteca TensorFlow. 33
3.3 Adicionando uma dimensão ao conjunto de entrada da rede. 34
3.4 Construção do LSTM através da biblioteca TensorFlow. 35

Lista de Tabelas

3.1 Atributos obtidos a partir do processo de integração dos dados. 31

4.1 Desempenho da rede neural do tipo MLP durante a etapa de Grid Search. . . . 38
4.2 Desempenho da rede neural do tipo LSTM durante a etapa de Grid Search. . . 39
4.3 Desempenho do Conjunto de Preditores no base de testes. 40
4.4 Desempenho de todos os modelos propostos. 41

Lista de Abreviaturas e Siglas

CCMC	Comunidade da NASA Centro de Modelagem Coordenada
GMU	lista de CME/ICME da George Mason *University*
GP	*Gaussian process*
GPR	*Gaussian Process Regressor*
GPR	*Gaussian Process Regressor*
GPR	*Gaussian Process Regressor*
LSTM	Long Short-Term Memory
LSTM	*Long Short-Term Memory*
MLP	*Multi-Layer Perceptron*
MPA	*Main Position Angle*
RC	Lista de Richardson e Cane
RNA	rede neural artificial
SDV	Synthetic Data Vault
SVR	*Support Vector Regression*
SVR	*Support Vector Regression*
SVR	*Support Vector Regression*
SVR	*support vector regression*
USTC	Universidade de Ciência e Tecnologia da China

Sumário

1 Introdução **12**
- 1.1 Hipótese . . . 13
- 1.2 Objetivos . . . 13
- 1.3 Organização do trabalho . . . 14

2 Levantamento bibliográfico **15**
- 2.1 Clima Espacial . . . 15
  - 2.1.1 Explosões Solares . . . 17
  - 2.1.2 Partículas Energéticas Solares (SEPs) . . . 17
  - 2.1.3 Ejeção de Massa Coronal (CME) . . . 17
  - 2.1.4 Consequências do Clima Espacial . . . 19
- 2.2 Séries Temporais . . . 21
- 2.3 Aprendizado de Máquina . . . 21
  - 2.3.1 *Support Vector Regression* . . . 22
  - 2.3.2 *Gaussian Process Regression* . . . 22
  - 2.3.3 *XGBoost* . . . 22
  - 2.3.4 Redes Neurais Artificiais . . . 23
  - 2.3.5 Otimização . . . 25
- 2.4 Pesquisas sobre o Tópico Estudado . . . 26
- 2.5 Considerações Finais . . . 28

3 Desenvolvimento **29**
- 3.1 Conjuntos de Dados . . . 29
  - 3.1.1 Integração dos dados . . . 30
  - 3.1.2 Geração de Dados Sintéticos . . . 32
- 3.2 Preditores . . . 32
  - 3.2.1 *Multi-Layer Perceptron* . . . 33
  - 3.2.2 *Long Short-Term Memory* . . . 34
  - 3.2.3 Preditor Conjunto . . . 35
  - 3.2.4 Otimização . . . 36
  - 3.2.5 Avaliação . . . 36

4 Resultados **37**
- 4.1 *Multi-Layer Perceptron* . . . 37
- 4.2 *Long Short-Term Memory* . . . 38
- 4.3 Conjunto de Preditores . . . 40
- 4.4 Considerações Finais . . . 40

5 Conclusões **42**

Capítulo 1

Introdução

Nas últimas décadas, muitos avanços tecnológicos ocorreram e a humanidade tornou-se mais dependente dos sistemas espaciais e serviços baseados em satélite, e de outras tecnologias influenciadas por fenômenos da interação Sol e Terra. A região onde as interações eletromagnéticas Sol-Terra são mais importantes é chamada de Geoespaço e abrange a atmosfera superior da Terra, a magnetosfera e os ventos solares locais (ECHER et al., 2005). O objetivo da pesquisa em clima espacial é entender o ambiente Geoespacial para viabilizar procedimentos que o torne seguro para as atividades tecnológicas e para a presença humana no espaço.

Dentre os tópicos pesquisados na área de Clima Espacial, estão as Ejeções de Massa Coronal (CMEs, do inglês), que são fenômenos violentos que ocorrem na atmosfera solar e acabam ecoando por toda a heliosfera. Elas consistem na expulsão de quantidades consideráveis de plasma das regiões da coroa do Sol. As CMEs são perigosas quando chegam na Terra, pois, ao colidir com a nossa magnetosfera, causam uma pertubação em todo tipo de aparelho da era da informação (HOWARD, 2014).

A principal motivação para se estudar as ejeções de massa coronal decorre do entendimento de massas, velocidades e energias envolvidas no fenômeno. Essa é uma das principais atividades solares e, além disso, as ejeções ocorrem em uma região do Sol dominada pelo campo magnético, deixando sua pesquisa ainda mais interessante dado que podemos relacionar conceitos de ambas áreas de estudo.

Sua atividade é associada com certos fenômenos solares, como as explosões solares, as quais muitos estudos apontam uma relação significativa entre os dois processos (YOUSSEF, 2012) (YASHIRO; GOPALSWAMY, 2008). Dessa maneira, o estudo de ejeções de massa coronal

não é só do interesse desse campo em particular, mas, também, ajuda na compreensão de outros eventos que ocorrem na superfície solar.

1.1 Hipótese

Devido à grande importância das CMEs e seus choques interplanetários, a previsão de seus tempos de chegada na Terra é um dos principais objetivos dos vários centros nacionais de previsão espacial (ZHAO; DRYER, 2014). Atualmente, a previsão do seu tempo de chegada é realizado a partir da utilização de algoritmos de aprendizado de máquina, os quais estão sendo aprimorados para detectar as principais variáveis que interferem no resultado final.

Muitos desses modelos que estão sendo criados conseguem estimar razoavelmente quando uma CME entrará em contato com o campo magnético terrestre, evitando maiores danos ao notificar as autoridades responsáveis. Nossa hipótese é que há um modelo para predição que obtenha um erro médio absoluto de apenas 10 horas, resultado semelhante aos obtidos pelos preditores baseados em algoritmos de aprendizado de máquina atuais.

1.2 Objetivos

Nesta seção serão apresentados tanto os objetivos gerais quanto os específicos deste projeto.

Este trabalho visa realizar a previsão do tempo de trânsito, em horas, de uma CME até chegar ao planeta Terra, identificando quais parâmetros e modelos de aprendizado de máquina tem um melhor desempenho frente aos demais. Já, dentre os objetivos específicos estão:

- Estudar conceitos relacionados a aprendizado de máquina, CMEs, predição de séries temporais, além de outros conceitos pertinentes para o desenvolvimento do projeto;
- Montar uma base de dados contendo as séries de dados históricos que serão utilizadas nos algoritmos;
- Implementar os preditores do tempo de chegada de uma CME;
- Realizar experimentos que permitam avaliar o desempenho dos preditores implementados.

1.3 Organização do trabalho

A presente monografia está organizada da seguinte forma: o Capítulo 1 justica o propósito do projeto e apresenta seus objetivos; o Capítulo 2 aborda todo o conceito teórico necessário para o desenvolvimento do trabalho e apresenta trabalhos correlatos a esse; o Capítulo 3 detalha as metodologias que foram utilizadas em todas as etapas do estudo; o Capítulo 4 descreve e analisa os resultados obtidos por meio das atividades; por fim, no Capítulo 5 o trabalho é concluido e são apresentadas ideias para trabalhos futuros.

Capítulo 2

Levantamento bibliográfico

Nesse capítulo serão apresentados os conceitos de clima espacial e suas consequências, aprendizado de máquina, séries temporais, rede neural multi-layer perceptron (MLP), rede neural long short-term memory (LSTM) e pesquisas semelhantes na literatura e seus resultados.

2.1 Clima Espacial

O termo clima espacial refere-se às condições do Sol e do vento solar, magnetosfera, ionosfera e termosfera que podem influenciar o desempenho e a confiabilidade de sistemas tecnológicos baseados no solo. Essas condições podem afetar a vida e a saúde humana (SCHWENN, 2006).

O vento solar é uma explosão intensa de radiação proveniente da liberação de energia magnética associada a manchas solares (FLETCHER et al., 2011). Esses fenômenos são vistos como áreas brilhantes ao sol e podem durar de minutos a horas. As principais formas de monitorar os ventos solares são em raios-x e luz óptica (SEVERNY, 1964).

Por saber-se que o espaço interplanetário compreendido entre o Sol e a Terra não é um vácuo, mas preenchido pelo vento solar que flui do Sol em direção aos planetas, e sendo o Sol uma estrela ativa, eventos ocorridos na fotosfera e na coroa solar causam alterações nas propriedades do vento solar cotidiano, aumentando sua densidade de partículas e sua velocidade, o que causa distúrbios na magnetosfera da Terra (KANE et al., 2008).

Segundo Eddy (1976), observações mostraram que o ciclo solar segue um padrão de aproximadamente 11 anos entre dois períodos de máxima atividade solar. Manchas solares, especialmente quando coexistem em grupos, podem liberar enormes quantidades de plasma

em alta velocidade arremessadas da superfície do Sol por um processo denominado reconexão magnética (GOLDSTON, 2020).

O número de Wolf (Rz) é uma medida da atividade solar diária que leva em conta a quantidade de manchas individuais e grupos de manchas na superfície solar (CLETTE et al., 2007). Outra forma de monitorar isso é utilizando a abordagem de Hoyt e Schatten (1998), o número de manchas solares de grupo (Rg), a qual acompanha a atividade solar de forma mais uniforme no longo prazo.

A Figura 2.1 mostra a média anual do número de Wolf (Rz), que está disponível desde 1700, e o número de manchas solares de grupo (Rg) que está disponível desde 1610. Podemos perceber a oscilação de 11 anos, bem como as variações de longo prazo.

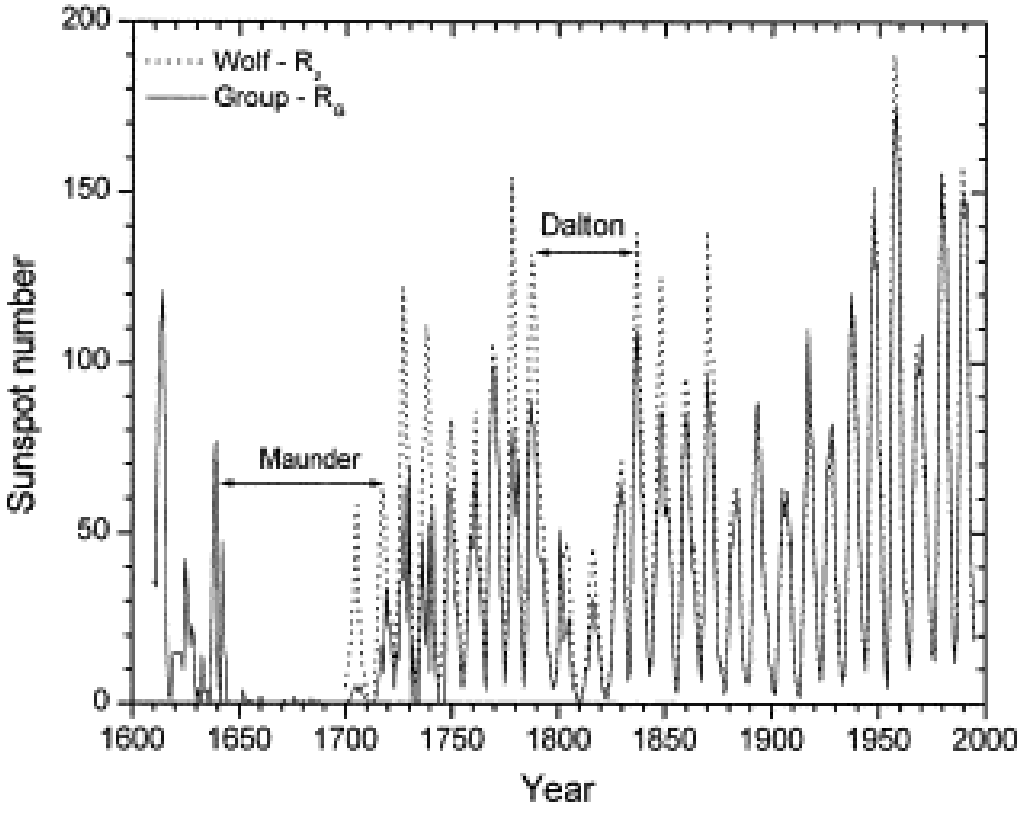

Figura 2.1: Média anual de manchas solares do tipo Wolf e de grupo entre os anos 1610-2000. Disponível em (EDDY, 1976).

Nossa sociedade tornou-se cada vez mais vulnerável aos distúrbios externos ao sistema da Terra, em particular aqueles iniciados por eventos explosivos no Sol, como:

1. Explosões solares, liberam picos de radiação cobrindo uma imensa faixa de comprimento de onda que podem aquecer a atmosfera terrestre em minutos.

2. Partículas Energéticas Solares (SEPs), aceleradas a energias quase relativísticas durante grandes tempestades solares, chegam à órbita da Terra em minutos e podem, entre outras situações, colocar em perigo os astronautas que viajam pelo espaço interplanetário.

3. Ejeções de Massa Coronal (CMEs), ejetadas no espaço interplanetário como nuvens gigantes de gás ionizado, que após algumas horas ou dias podem atingir a Terra.

2.1.1 Explosões Solares

Carrington (1859) relatou pela primeira vez a ocorrência de uma erupção solar, uma manifestação vista quando ele observou manchas solares através de um telescópio. Na época, concluiu-se a partir dessa observação casual que a perturbação da atmosfera solar foi compacta, breve e extremamente energética. A ideia de que uma perturbação solar poderia afetar um instrumento terrestre, como uma bússola, parecia altamente improvável na época, mas acabou sendo uma associação correta. Então essa primeira erupção observada serviu para sugerir a capacidade de tal evento solar ter influências amplamente sentidas por nós.

Explosões solares podem liberar energia equivalente a bilhões de bombas atômicas em poucos minutos. Elas liberam rajadas de raios-X, partículas carregadas, que podem mais tarde atingir a Terra, colocando em perigo os satélites e causando apagões de energia.

2.1.2 Partículas Energéticas Solares (SEPs)

Os eventos SEP são observados como elétrons, prótons e íons pesados excedendo os níveis de energia térmica. Esses eventos são controlados por muitos fatores, como a localização da região de origem da erupção (longitude e latitude) e largura da CME, vento solar e campo magnético próximo ao choque.

Fortes fluxos de energia protoneladas (os chamados eventos de prótons) causam fenômenos geoefetivos mais fortes. SEPs de alta energia atingem a Terra em menos de 10 minutos e causam benefícios, como a melhoria no nível do solo, mas também prejuízos, como o aumento da exposição à radiação.

2.1.3 Ejeção de Massa Coronal (CME)

Embora as pesquisas desenvolvidas em clima espacial possam abranger uma ampla gama de efeitos, uma de suas principais preocupações é a interação de Ejeções de Massa Coronal

(CMEs) com a magnetosfera, ionosfera, atmosfera e litosfera da Terra. A força dessa interação é controlada em grande parte pela velocidade da CME que chega e pela quantidade de campo magnético apontando para o sul contido nela.

CMEs são caracterizadas por velocidade, largura angular e um ângulo de posição central no plano do céu. As velocidades frontais medidas variam de alguns km/s (perto do Sol) a quase 3.000 km/s (GOPALASWAMY; NARAYANAN, P.; NARAYANAN, S., 2004).

Um enorme catálogo com mais de 10.000 CMEs foi reunido por Yashiro, Gopalswamy et al. (2004), todos descobertos após inspeção visual dos dados dos instrumentos SOHO. Algumas informações importantes, por exemplo, imagens tiradas pelos instrumentos SOHO, animações, imagens de diferença e algumas propriedades características estão presentes nesse catálogo. Na figura 2.2 podemos ver uma das diversas imagens capturadas pelo SOHO. A fim de torná-lo facilmente acessível à comunidade científica, é possível acessá-lo através de um site específico [1].

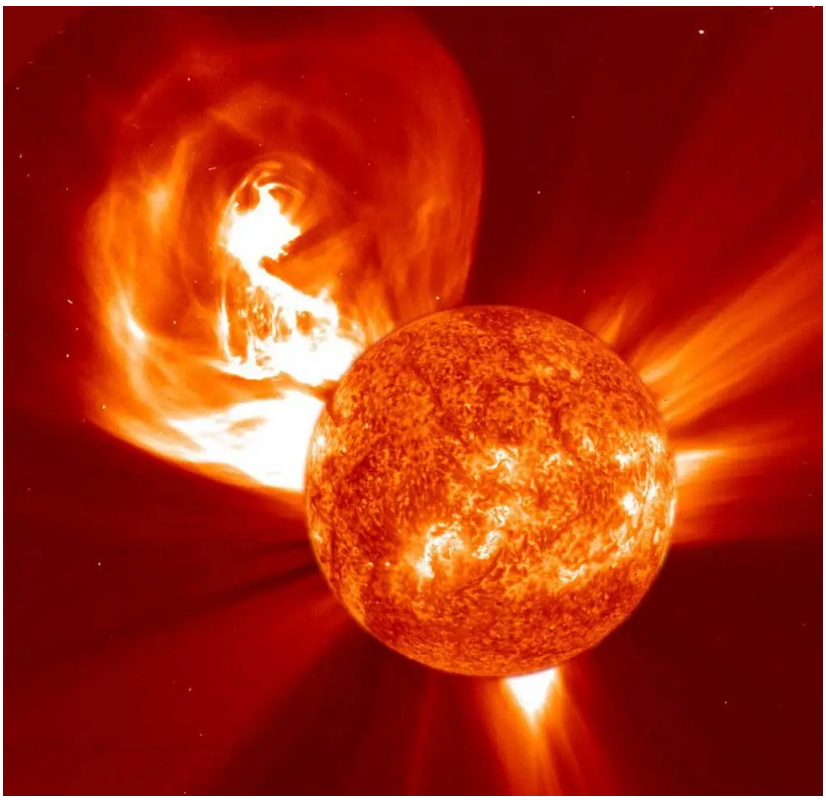

Figura 2.2: Ejeção de massa coronal pelo Sol. Disponível nesse site específico [1]

O início das CMEs é descrito em um cenário de três fases por Jie Zhang et al. (2001): a fase de iniciação, a fase de aceleração impulsiva e a fase de propagação. A fase de iniciação

[1] O site está disponível em `http://cdaw.gsfc.nasa.gov/CME_list`.

sempre ocorre antes do início de um fenômeno associado e leva algumas dezenas de minutos. Já a fase de aceleração impulsiva coincide com a ascensão da explosão, parando apenas com o pico de explosões de raios-X suaves. As CMEs então passam por uma fase de propagação, que é caracterizada por uma velocidade constante da CME.

2.1.4 Consequências do Clima Espacial

Como ficou cada vez mais claro ao longo do tempo, os satélites modernos, as instalações meteorológicas e de comunicação global, e muitos sistemas de defesa tornaram-se dependentes do conhecimento da condição do ambiente espacial próximo da Terra (BAKER, 1998).

As consequências econômicas desses efeitos são enormes, tanto que essa é uma das razões pelas quais o clima espacial e sua previsibilidade chamam grande atenção recentemente. Outra razão, é a nova série de dados observacionais obtidos na última década a partir de instrumentos espaciais não disponíveis antes, o que melhora a qualidade dos dados para futuras pesquisas no tema.

Apesar dos esforços na área, a previsão dos efeitos do clima espacial ainda continua sendo um grande desafio. A precisão e confiabilidade na previsão de eventos espaciais e seus impactos são muito fracos. Eles ocorrem de forma bastante espontânea, e até o momento não se identificou assinaturas únicas que indicariam um fenômeno e seu provável tempo de início, localização e impacto na Terra.

A seguir, são apresentados brevemente os efeitos do clima espacial nas mais diversas áreas da nossa sociedade.

Redes Elétricas

As correntes induzidas geomagneticamente (GICs) associadas a tempestades geomagnéticas podem danificar a infraestrutura física (especificamente transformadores), introduzir instabilidades de tensão que podem levar a um apagão sem danificar a infraestrutura e interferir nos sistemas de proteção e detecção de falhas. De acordo com (TSURUTANI, 2003), é importante notar que os sistemas de correntes ionosféricas que se acoplam aos GICs são muito estruturados e são mais intensos em latitudes relativamente altas nas proximidades das ovais aurorais.

Indústria do Petróleo e Gás

Os GICs podem causar mudanças na tensão do tubo de petróleo e gás que está no solo, o que acaba por aumentar a corrosão (VILJANEN et al., 2006). De acordo com Viljanen et al. (2006), pesquisas aeromagnéticas e perfuração de precisão são afetadas por flutuações magnéticas que ocorrem durante as tempestades geomagnéticas.

Comunicações

Certas redes móveis podem ser afetadas pela perda de informações de tempo do sistema global de navegação por satélite (GNSS), que é usado pela aviação, navegação e militares (KELLY et al., 2014) (CANNON, 2013). Durante tempestades geomagnéticas, ocorrem reduções regionais e globais na banda operacional de alta frequência. Os sistemas de alta frequência modernos são projetados para serem resilientes à essas falhas, mas sistemas legados podem sofrer interrupções. Por fim, as redes de fibra ótica exigem estações repetidoras para aumentar periodicamente o sinal, portanto a infraestrutura de energia dessas estações está em risco quando há GICs (LANZEROTTI et al., 2001).

Satélites

Segundo Wrenn, Rodgers e Ryden (2002), elétrons energéticos presos no cinturão de radiação externo causam carga e descarga eletrostática, o que pode danificar equipamentos eletrônicos sensíveis e painéis solares. Durante as tempestades de Halloween de 2003, 47 satélites relataram anomalias (aproximadamente 10% dos satélites em órbita), um satélite científico foi perdido e 10 satélites perderam serviço operacional por mais de um dia (CANNON, 2013) (BARBIERI; MAHMOT, 2004).

Aviação

Tempestades de radiação solar aumentam o ambiente de radiação gerada por raios cósmicos na altitude de voo (DYER et al., 2003). Um efeito talvez contra-intuitivo é que a radiação de partículas energéticas pode diminuir durante ou depois de uma tempestade geomagnética (GETLEY et al., 2005). Devido ao clima espacial severo, pode ser necessário reduzir o tempo de voo em grandes altitudes (DYER et al., 2003), e isso teria um impacto comercial e operacional, incluindo atrasos e aumento do uso de combustível (EASTWOOD et al., 2017), já que os eventos chegam sem aviso e podem persistir por várias horas. Uma perda severa de rádio de alta

frequência pode levar à perda de comunicações com a maioria das aeronaves no Atlântico Norte. Nessa situação, as aeronaves já em voo continuariam, mas as que estavam no solo provavelmente não teriam permissão para decolar (CANNON, 2013).

2.2 Séries Temporais

Uma série temporal representa uma coleção de observações feitas de maneira sequencial em um período. A previsão por meio de séries temporais é uma prática amplamente adotada, na qual é possível estimar o valor futuro de uma variável tendo como base o seu histórico (ADHIKARI; AGRAWAL, 2013). Essas estimativas são de extrema útilidade pois apoiam o nosso processo de tomada de decisão.

Uma série temporal é definida por suas características, as quais afetam em sua análise e modelagem. Dentre elas, a estacionariedade de uma série é medida pela utilização da média e variância, caso elas sejam iguais ou próximas durante todo tempo observado, a série é estacionaria. Já a sazonalidade, identifica uma repetição de comportamento durante um período específico. Um exemplo seria o consumo de energia no inverno, que sofre um acréscimo na quantidade gasta dado que as pessoas tendem a utilizar mais energia durante esse período específico do ano.

A análise de uma série temporal têm o objetivo de modelar o fenômeno em questão, obter conclusões em termos estatísticos e avaliar a adequação do modelo em termos de previsão (CASTRO, 2001).

2.3 Aprendizado de Máquina

O aprendizado de máquina é uma abordagem interessante para predição do tempo de chegada de uma CME, porque ele permite a inferência de relacionamentos de dados que muitas vezes não são diretamente observáveis por humanos (CAMPOREALE, 2019). As técnicas de aprendizado de máquina são caracterizadas por investigar como as máquinas podem adquirir conhecimento através da extração de padrões a partir de um conjunto de dados, buscando o desenvolvimento de algoritmos que permitam que computadores possam se tornar capazes de tomar decisões com certa autonomia (XU et al., 2021).

2.3.1 *Support Vector Regression*

O *Support Vector Regression* (SVR) é um algoritmo de aprendizado de máquina usado para análise de regressão. Seu objetivo é encontrar uma função que aproxime a relação entre variáveis de entrada a uma variável de saída enquanto minimiza o erro da predição (AWAD et al., 2015). O SVR é generalizado pelo *Support Vector Machine* (SVM), um modelo de aprendizado de máquina usado para classificação de dados. O SVM é um dos métodos de predição mais robustos, sendo baseado em estruturas de aprendizado estatístico. Já o SVR é considerado superior e suporta regressões lineares e não lineares (DRUCKER et al., 1996).

Ao contrário dos métodos de regressão tradicionais que dependem de suposições do modelo que podem não ser precisas, o SVR é uma técnica de aprendizado de máquina na qual um modelo aprende a importância de uma variável a fim de caracterizar a relação entre entrada e saída (ZHANG, F.; O'DONNELL, 2020).

2.3.2 *Gaussian Process Regression*

É uma abordagem bayesiana não paramétrica para regressão que nos permite fazer predições sobre nossos dados incorporando conhecimento prévio. O *Gaussian Process Regression* (GPR) funciona bem em pequenos conjuntos de dados e tem a capacidade de fornecer medições de incerteza em suas predições.

Os processos gaussianos são um método genérico de aprendizado supervisionado projetado para resolver problemas de regressão e classificação probabilística. A previsão é probabilística (Gaussiana) para que se possa calcular intervalos de confiança e decidir com base neles se deve-se reajustar a predição.

2.3.3 *XGBoost*

O *XGBoost* é um algoritmo de aprendizado de máquina baseado em árvore de decisão que usa uma estrutura de aumento de gradiente (CHEN, T.; HE et al., 2015). Sua nomenclatura vem do termo *Extreme Gradient Boosting*, ele tornou-se um dos algoritmos de aprendizado de máquina mais populares e amplamente utilizados atualmente devido à sua capacidade de lidar com grandes conjuntos de dados e de obter desempenho de ponta em muitas tarefas de aprendizado de máquina, como classificação e regressão.

Um dos principais recursos do XGBoost é a manipulação eficiente de valores ausentes, sem a necessidade de pré-processamento significativo (CHEN, T.; HE et al., 2015). Esse algoritmo é altamente personalizável, permitindo o ajuste de vários parâmetros do modelo para otimizar o seu desempenho (MITCHELL; FRANK, 2017).

2.3.4 Redes Neurais Artificiais

A rede neural artificial (RNA) é um componente capaz de processar informação de forma paralela com a capacidade de armazenar conhecimento e experiência a fim de torná-lo disponível para uso (KUBAT, 1999). A RNA é formada por neurônios artificiais cujo objetivo, a exemplo dos neurônios do cérebro humano, é aprender e tomar decisões baseadas em seu próprio aprendizado (FLECK et al., 2016).

O Perceptron é um modelo de neurônio criado por Rosenblatt (1958) para reconhecimento de padrões. Podemos observar o funcionamento do Perceptron na figura 2.3. Primeiramente, aplicamos pesos (W) nos valores de entrada (X) e na constante, que está servindo como viés. Em seguida, realiza-se um somatório ($\sum$) dos resultados obtidos. Após isso, passamos pela função de ativação a fim de introduzir um componente não-linear no Perceptron, fazendo com que eles possam aprender mais do que relações lineares entre as variáveis (SHARMA; SHARMA; ATHAIYA, 2017). Por fim, o resultado obtido após a aplicação da função de ativação é o valor de saída do Perceptron.

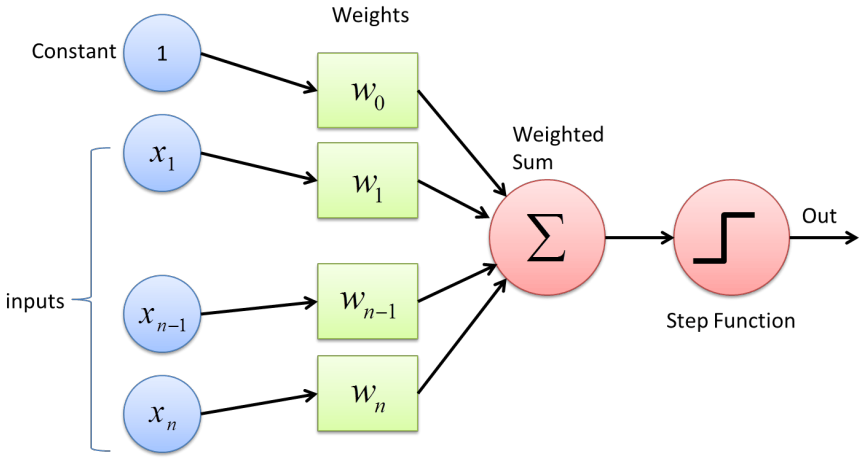

Figura 2.3: Funcionamento do Perceptron. Disponível em Ramchoun et al. (2016).

Dentre todas as funções de ativação, a função ReLU é uma das mais utilizadas nos trabalhos envolvendo RNAs. Seu acrônimo significa Rectified Linear Unit (Unidade Linear Retificada, em português). Seu funcionamento ocorre da seguinte maneira: o valor de saída é o valor máximo

entre 0 e x, assumindo zero quando x é menor que zero e x quando esse valor é maior do que zero (HAHNLOSER et al., 2000). Foi demonstrado que essa função permite um melhor treinamento de Redes Neurais, gerando menor erro durante o treinamento quando comparada com as outras funções de ativações (HARA; SAITO; SHOUNO, 2015).

Uma rede neural deve ser configurada de modo que um conjunto de entradas produz o conjunto desejado de saídas. Uma maneira de definir os valores dos pesos é treinar a rede neural e deixá-la mudar seus próprios pesos de acordo com alguma regra de aprendizado. A aprendizagem em redes neurais podem ser classificadas em três gêneros distintos: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, um vetor de entrada é dado como entrada junto com um conjunto de saídas desejadas. Após a primeira passagem do conjunto de entrada pela rede, o que é chamado de iteração, realiza-se uma comparação entre a saída obtida e a saída esperada. Os resultados dessa comparação são então usados para determinar as mudanças de peso na rede de acordo com a regra de aprendizado vigente. Isso é feito repetidamente até alcançarmos um erro aceitável ou atingirmos o número de iterações pré-estabelecida.

Multi-Layer Perceptron

Dentre as técnicas de aprendizado de máquina, encontram-se as redes neurais artificiais, que têm sido amplamente utilizadas para problemas de regressão, como é o caso da predição de séries temporais. A maioria das pesquisas voltadas à predição de dados recorre à rede neural do tipo Multi-Layer Perceptron (MLP), consideradas aproximadoras universais de funções, para realizar este tipo de tarefa.

A MLP consiste em um sistema de nós interconectados. Os nós são conectados por pesos e sinais de saída que são uma função da soma das entradas para o nó modificado por uma transferência não linear simples, ou função de ativação (GARDNER; DORLING, 1998). A arquitetura de um perceptron multicamada é variável, mas em geral consistirá em várias camadas de neurônios. A camada de entrada não desempenha nenhum papel computacional, mas apenas serve para passar o vetor de entrada para a rede (MURTAGH, 1991).

Aprender para o MLP é o processo de adaptar as conexões pesos a fim de obter uma diferença mínima entre a saída da rede e a saída desejada, por isso na literatura alguns algoritmo são usados. Deles, o mais usado é o de retropropagação, o qual é baseado em técnicas de gradiente descendente (SALOMON, 2004).

Long Short-Term Memory

Outra técnica muito popular na literatura e também utilizada nesse trabalho é a *Long Short-Term Memory* (LSTM). A arquitetura LSTM consiste em um conjunto de sub-redes conectadas recorrentemente, conhecidos como blocos de memória. Esses blocos podem ser pensados como um bloco diferenciável versão dos chips de memória em um computador digital (GRAVES; GRAVES, 2012). Cada bloco contém uma ou mais células de memória autoconectadas e três unidades multiplicativas – os portões de entrada, saída e esquecimento – que fornecem análogos de gravação, leitura e redefinição das operações para as células (HOCHREITER; SCHMIDHUBER, 1997).

Uma rede LSTM é igual a uma rede neural artificial padrão, exceto que as unidades de soma na camada oculta são substituídas por blocos de memória. O modelo LSTM é treinado usando o gradiente completo, conforme apresentado por Graves e Schmidhuber (2005) para ajustar os pesos envolvidos na rede.

2.3.5 Otimização

Algoritmos de aprendizado de máquina têm sido amplamente utilizados em diversas aplicações e áreas. Para ajustar um modelo de aprendizado de máquina em diferentes problemas, seus hiperparâmetros devem ser ajustados. Selecionar a melhor configuração de hiperparâmetros para modelos de aprendizado de máquina tem um impacto direto no desempenho do modelo (YANG; SHAMI, 2020). Embora existam várias técnicas de otimização, elas têm diferentes vantagens e desvantagens quando aplicadas a diferentes tipos de problemas (GAMBELLA; GHADDAR; NAOUM-SAWAYA, 2021).

Validação Cruzada com Janela Deslizante

Janela deslizante é uma aproximação temporária sobre o valor real da série temporal dados. O tamanho da janela e do segmento aumenta até atingirmos o menor erro aproximação (MOZAFFARI, L.; MOZAFFARI, A.; AZAD, 2015). Após selecionar o primeiro segmento, o próximo segmento é selecionado a partir do final do primeiro segmento. O processo é repetido até que todos os dados da série temporal sejam segmentados. O processo de janela deslizante é mostrado na Figura 2.4 com tamanho de janela igual a 5.

A janela deslizante acumula os dados históricos da série temporal para prever o dia seguinte (YU et al., 2014). A Figura 2.5 mostra o processo de janela deslizante com tamanho de janela

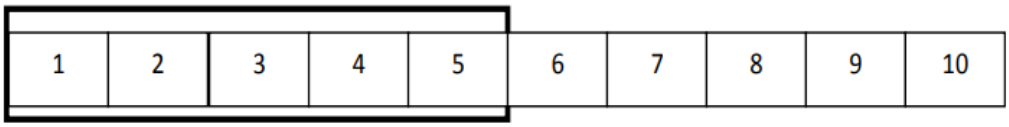

Figura 2.4: Estado inicial da janela deslizante.

igual a 5. Cada número $(1, 2, 3, \ldots, 10)$ representa um evento CME. Inicialmente, a janela cobriu de 1 a 5, representando dados históricos de 5 CMEs que estão sendo usados para previsão do tempo de trânsito da próxima CME na base. Então, a janela desliza para o lado direito por uma observação para cobrir outras 5 observações (de 2 a 6) para prever o tempo de trânsito da próxima CME. O processo será continuado até os dados da série temporal se esgotarem.

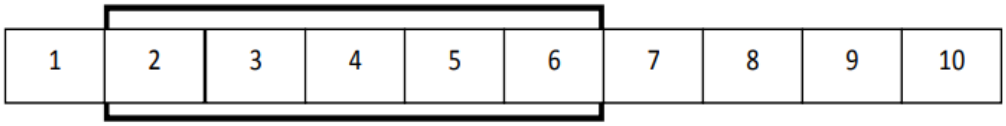

Figura 2.5: Estado da janela após deslizar.

Grid Search

A busca em grade é originalmente uma pesquisa exaustiva baseada em um subconjunto definido do espaço de hiperparâmetros. O desempenho de cada combinação é avaliado usando algumas métricas de desempenho. A busca em grade otimizou os parâmetros das redes neurais do tipo MLP e LSTM usando a técnica de validação cruzada da Seção 2.3.5 como uma métrica de desempenho. O objetivo é identificar uma boa combinação de hiperparâmetros para que o preditor possa prever dados desconhecidos com precisão. A combinação com a menor média das métricas definidas na Seção 3.2.5 é selecionada e usada para treinar todo o conjunto de dados.

2.4 Pesquisas sobre o Tópico Estudado

Existem várias formas em que podemos abordar o problema da predição do tempo de trânsito de uma CME até a Terra, passando por modelos baseados na Física e modelos criados por algoritmos de aprendizado de máquina. O objetivo desta seção é apresentar ao leitor os principais modelos construídos até os dias de hoje juntamente com os seus resultados.

No modelo *Shock Time of Arrival* (STOA) de Dryer (1974), é considerado que a onda de choque possui velocidade constante por um intervalo de tempo definido. Esse intervalo é dado pela duração da emissão de raios-x associada ao evento em um vento solar com velocidade de 1 Unidade Astronômica (UA). Depois da condução, a velocidade do choque vai diminuindo, com uma proporção igual a R-1/2 (onde R é a distância do Sol) (PARKER, 1963). Isso e mais um choque presumido determinam quando a CME chegará à Terra. A energia que o choque inicialmente adquire durante seu lançamento prescrito (velocidade mais duração) junto com a forma de choque assumida e as condições de vento solar assumidas determinam o quão rápido o choque enfraquece e com que força ele chegará na Terra. Fry et al. (2003) aplicou o modelo STOA para 173 eventos a fim de compilar informações sobre seu desempenho. A raiz do erro quadrático médio nos tempos previstos de chegada da CME foi de 12,2 h.

Enquanto isso, Zhao e Dryer (2014) forneceram um exame extensivo das habilidades de previsão e precisão de vários modelos de previsão de tempo de trânsito CME, incluindo os baseados em física ((FENG et al., 2009; MILLWARD et al., 2013)), empíricos ((KIM; MOON; CHO, K.-S., 2007; MICHALEK; PUCHOWSKA; RAMS, 2009)) e modelos baseados em deslocamento (SONG, 2010; VRŠNAK et al., 2013)). Eles obtiveram um erro médio absoluto de 10 h para quase todos os modelos considerados em seu estudo.

Recentemente, algoritmos de aprendizado de máquina foram aplicadas de forma contínua e eficaz a vários desafios de previsão do clima espacial para produzir resultados promissores. Em vez de um simples ajuste de curva, uma rede neural bem projetada pode melhorar um modelo teórico por meio da generalização. Wang et al. (2019) iniciaram a aplicação da CNN com imagens de luz branca do LASCO como entrada para prever o tempo de chegada de 223 CMEs e obtiveram um erro médio absoluto de 12,4 h. Sudar, Vršnak e Dumbović (2015) adotaram uma rede neural para calcular o tempo de trânsito de 153 CMEs com as velocidades iniciais dos CMEs como entradas. A previsão desviou-se do tempo de observação em uma média de 12 h. A série de modelos foi seguida por uma CNN com um resultado significativamente melhor (erro médio absoluto de 5,8 h) por Fu et al. (2021), o qual utilizou 2.400 imagens de CMEs como entrada do modelo para prever o tempo de chegada de CMEs.

2.5 Considerações Finais

Neste capítulo foi abordada a fundamentação teórica sobre o clima espacial e suas consequências, as séries temporais e o aprendizado de máquina, além de trabalhos correlatos na literatura.

Foram apresentados os principais fenômenos e consequências do clima espacial, incluindo a CME, que é o foco desse trabalho. Dessa maneira, podemos perceber o tamanho do impacto deles na nossa sociedade ao danificarem aparelhos da era digital.

Para o aprendizado de máquina, abordamos sua utilidade para o nosso problema. Mais especificamente, foram descritas as características e principais propriedades dos modelos MLP e LSTM, que foram empregados nesse trabalho.

Por fim, citamos os principais modelos de predição do tempo de trânsito de uma CME até a Terra e apresentamos os seus resultados. Mostramos ao leitor desde modelos baseados em aprendizado de máquina até modelos baseados em conceitos físicos.

Capítulo 3

Desenvolvimento

Este capítulo apresenta uma contextualização sobre os dados e a metodologia utilizada. Além disso descrevemos o estudo e avaliação da série temporal empregada no projeto e realizamos ponderações importantes.

A análise realizada, desde a coleta dos dados até a realização de modelagem e teste, foi feita utilizando a linguagem Python, versão 3, por meio de máquinas gratuitas disponibilizadas pelo ambiente do Google Colab. Foram utilizadas as bibliotecas Pandas (MCKINNEY et al., 2011); NumPy (OLIPHANT et al., 2006) e datetime para tratamento dos dados; Keras (CHOLLET et al., 2015) e TensorFlow (ABADI et al., 2015) para criação de instâncias dos modelos de rede neural MLP e LSTM; a biblioteca Scikit-Learn (PEDREGOSA et al., 2011) para instâncias dos modelos SVR e GPR; e a biblioteca XGBoost (CHEN, T.; GUESTRIN, 2016b) para instâncias do modelo XGBoost. Além disso, usamos a biblioteca Synthetic Data Vault (SDV) (PATKI; WEDGE; VEERAMACHANENI, 2016) para geração e adição de dados sintéticos à base de dados que utilizamos 3.1.2.

3.1 Conjuntos de Dados

Seguindo a recomendação de Jiajia Liu et al. (2018), adotamos quatro listas de CME: 1) a Lista de Richardson e Cane (RC) (RICHARDSON; CANE, 2010)[1]; 2) a lista completa de CME mantida pela Universidade de Ciência e Tecnologia da China (USTC) (SHEN et al., 2014)[2]; 3) a lista

[1]Disponível em `http://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm`.
[2]Disponível em `http://space.ustc.edu.cn/dreams/fhcmes/index.php`.

de CME/ICME da George Mason *University* (GMU) (HESS; ZHANG, J., 2017)[3]; e 4) o CME *Scoreboard* mantido pela Comunidade da NASA Centro de Modelagem Coordenada (CCMC)[4].

Combinamos os CMEs das quatro listas com base nos procedimentos descritos no trabalho de Jiajia Liu et al. (2018). Esse processo resultou em um conjunto de 363 CMEs de 1996 a 2021. Para cada uma das 363 CMEs reportadas, coletamos 12 atributos relacionados a CME, energia solar e vento solar. A seguir, descrevemos nossos dados e o processo de integração.

3.1.1 Integração dos dados

Para cada lista CME, coletamos os horários de início e chegada dos CMEs. Em seguida, combinamos as listas em uma única base de dados, removendo duplicatas ao examinar por eventos que ocorreram no mesmo momento no conjunto de dados. Conforme os procedimentos descritos em Jiajia Liu et al. (2018), foram obtidas 216 CMEs do RC, 24 do USTC, 38 do GMU e 113 do CCMC.

Além das quatro listas, também consideramos o catálogo apresentado em Paouris e Mavromichalaki (2017), que continha 266 eventos CME. Juntamos esses eventos com os obtidos anteriormente, mantendo apenas um evento entre os ocorridos na mesma hora. Tal integração de dados resultou em 363 amostras em nossa base de dados.

Para cada evento em nosso conjunto de dados, obtivemos suas cinco características do Catálogo SOHO LASCO CME[5]. As características do CME são a largura angular, o ângulo de posição principal (*Main Position Angle* – MPA), a velocidade linear, a velocidade de segunda ordem na altura final e a massa. Especificamente, para cada evento em nosso conjunto de dados, usando seu tempo de início como índice, conseguimos suas cinco características no Catálogo LASCO CME e integramos os atributos com o horário de chegada do evento em nossa base de dados. Para um evento E na base cujo início não corresponde ao início de qualquer CME no Catálogo LASCO CME, selecionamos os cinco atributos temporalmente mais próximos de E, dentro do intervalo de uma hora, e atribuímos elas ao evento não correspondente com o início de uma CME no catálogo.

No Catálogo LASCO CME, um atributo A pode ter um valor ausente. Portanto, esse valor foi transferido para nossa base de dados. Nós empregamos uma técnica de limpeza de dados

[3]Disponível em `http://solar.gmu.edu/heliophysics/index.php/GMU_CME/ICME_List`.
[4]Disponível em `https://kauai.ccmc.gsfc.nasa.gov/CMEscoreboard`
[5]Disponível em `https://cdaw.gsfc.nasa.gov/CME_list/index.html`.

que consiste em calcular a média dos valores disponíveis para o atributo A em nosso conjunto de dados e usamos a média para representar o valor ausente.

Para obter as características do vento solar de cada evento em nosso conjunto de dados, seguimos a abordagem descrita em Jiajia Liu et al. (2018) e utilizamos os dados extraídos do NASA OMNIWeb[6]. Nós consideramos sete parâmetros do vento solar: a relação alfa para próton, a longitude do fluxo, a pressão do plasma, Bx, Bz, a velocidade do fluxo e a temperatura dos prótons. Para cada evento E em nosso conjunto de dados, usamos sua hora de início t como índice, obtivemos os atributos do vento solar no tempo $t + 6$ caso o início não correspondesse com nenhuma CME do Catálogo LASCO CME, e atribuímos esses valores ao evento E.

O resultado final do processo de integração originou uma nova base no formato 363×14, contendo dados sobre a CME e sobre o vento solar. A tabela 3.1 mostra cada atributo obtido e suas respectivas descrições.

Tabela 3.1: Atributos obtidos a partir do processo de integração dos dados.

Atributo	Descrição
disturbance	horário de início da tempestade geomagnética associada
transit_time	velocidade média da perturbação
angular_width	largura angular da CME
MPA	ângulo de posição principal da CME
avg_speed	velocidade média da CME
final_speed	velocidade final da CME
mass	massa da CME
Bx	eixo x do campo magnético interplanetário
Bz_GSE	eixo z do campo magnético interplanetário
plasma_temp	temperatura do plasma do vento solar
plasma_speed	velocidade do plasma do vento solar
plasma_flow_long	tamanho do fluxo de plasma do vento solar
alpha_prot_ratio	proporção da temperatura de partículas alpha e próton do vento solar
flow_pressure	pressão do fluxo de plasma do vento solar

[6]Disponível em `https://omniweb.gsfc.nasa.gov`.

3.1.2 Geração de Dados Sintéticos

Conforme indicamos na Seção 3.1.1, extraímos um total de 363 amostras da integração dos dados. Esse número de amostras é relativamente baixo e, dessa maneira, podemos classificar nosso cenário como um problema de *small data*, isto é, um conjunto pequeno de amostras para ter um bom treinamento dos algoritmos de aprendizado de máquina. Para minimizar os efeitos causados pela baixa quantidade de instâncias, adotamos a técnica de geração de dados sintéticos.

Esse método consiste em criar dados artificiais que imitam os dados reais ao invés de usar os produzidos por eventos que ocorreram de fato. Os dados são originados através de algoritmos e podem ser usados para treinamento de modelos de aprendizado de máquina e validação de modelos matemáticos.

Utilizamos a biblioteca SDV (PATKI; WEDGE; VEERAMACHANENI, 2016) para produzir os dados sintéticos em nossa base de dados. Ela consiste em um sistema que constrói modelos generativos a partir de bancos de dados relacionais. A biblioteca utiliza uma abordagem de modelagem multivariada para imitar as amostras, iterando através de todas as relações possíveis na base de dados.

Definimos que ela seria formada por 50% de dados reais e 50% de dados sintéticos tendo em vista aumentar o número de amostras, mas sem deixar que os dados reais tivessem uma participação pequena no conjunto de dados. Dessa maneira, conseguimos um total de 726 amostras em nossa base, sendo metade formada por dados sintéticos, destinados exclusivamente para o treinamento dos modelos de aprendizado.

3.2 Preditores

Esta seção apresenta os algoritmos de aprendizado de máquina empregados no nosso problema. Os preditores se utilizam dos dados em formato de séries temporais para realizar a predição do tempo de trânsito de uma CME. Desse modo, as redes neurais do tipo *Multi-Layer Perceptron* (MLP) e *Long Short-Term Memory* (LSTM) foram consideradas devido a sua capacidade de lidar com problemas de regressão. Além delas, foi implementado um preditor que leva em sua arquitetura outros três algoritmos de aprendizado de máquina.São eles o *Support Vector Regression* (SVR), o *Gaussian Process Regressor* (GPR) e o XGB *Regressor*.

3.2.1 *Multi-Layer Perceptron*

Uma rede neural artificial (RNA) do tipo MLP foi implementada com o auxílio das bibliotecas Tensorflow (ABADI et al., 2015) e Keras (CHOLLET et al., 2015). Tensorflow é uma biblioteca para Python que reúne diversos algoritmos de aprendizado de máquina de última geração para lidar com problemas supervisionados e não supervisionados. Ela possui um foco particular no treinamento e inferência de redes neurais artificiais. O Keras é outro pacote disponível na linguagem Python, ele é a interface do Tensorflow para criar e treinar modelos de aprendizado profundo. Essa biblioteca também pode ser usada de maneira independente do Tensorflow.

A primeira etapa para o processo de predição foi a seleção de atributos da nossa base de dados. Para isso, selecionamos como entrada para o modelo MLP apenas os atributos `angular_width`, MPA, `avg_speed`, `final_speed` e BX e como saída os valores contidos em `transit_time`, conforme feito por Wang et al. (2019).

```
[ ]  desired_columns = ['angular_width', 'MPA', 'avg_speed', 'final_speed', 'BX', 'flow_pressure']
     target_columns = ['transit_time']

     feature = df[desired_columns]
     target = df[target_columns]
```

Figura 3.1: Seleção de atributos que servirão de entrada para os modelos.

```
[ ]  def MLP (hidden_layers, neurons, learning_rate):
       MLP_model = tf.keras.models.Sequential()

       for i in range(hidden_layers):
         layer = tf.keras.layers.Dense(neurons, activation='relu')
         MLP_model.add(layer)

       output_layer = tf.keras.layers.Dense(1, activation='relu')
       MLP_model.add(output_layer)

       MLP_model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=learning_rate),
                         loss='mse', metrics=['mean_absolute_error'])

       return MLP_model
```

Figura 3.2: Construção da MLP através da biblioteca TensorFlow.

Após isso, o MLP foi construído como mostra a Figura 3.2. A rede da Figura 3.2 foi criada como uma função para se adequar às próximas etapas da metodologia. Em todas as camadas foi utilizada a função de ativação reLU. Essa função é linear na sua dimensão positiva e zero na sua dimensão negativa, i.e. a sua saída é o valor máximo entre zero e o valor de entrada.

Definimos a reLU como função de ativação de todas camadas devido a sua simplicidade, amplo uso na literatura e, quando comparada com as funções sigmoide e tangente hiperbólica, menor incidência de problemas com o gradiente, como ele desaparecer ou estourar.

O otimizador Adam (KINGMA; BA, 2014) foi escolhido para o nosso modelo, levando como parâmetros o erro médio absoluto como métrica de avaliação. O otimizador também possui uma taxa de aprendizado a ser definida nos próximos passos do algoritmo. A taxa de aprendizado é um parâmetro do otimizador que determina o tamanho de ajuste dos pesos a cada iteração enquanto se move em direção a uma função de perda mínima. Ela afeta a velocidade na qual um modelo de aprendizado de máquina "aprende". Escolher a taxa de aprendizado é desafiador, pois um valor muito pequeno pode resultar em um longo processo de treinamento, enquanto um valor muito grande pode resultar no aprendizado de um conjunto de pesos abaixo do ideal.

3.2.2 *Long Short-Term Memory*

A Long Short-Term Memory (LSTM) também foi construída com o auxílio das bibliotecas Tensorflow e Keras. Para a implementação deste tipo de RNA fizemos a mesma seleção de atributos realizada para a MLP, como demonstrado na Seção 3.2.1. Além dessa etapa, realizamos uma transformação do nosso conjunto de entrada a fim de deixá-lo no formato aceito pela arquitetura da LSTM no Tensorflow. Dessa maneira, o conjunto passou a ter três dimensões, conforme descrito na Figura 3.3.

```
[ ]  X_train = X_train.values.reshape(X_train.shape[0], 1, X_train.shape[1])
     X_test = X_test.values.reshape(X_test.shape[0], 1, X_test.shape[1])

     X_train.shape
     X_test.shape
```

Figura 3.3: Adicionando uma dimensão ao conjunto de entrada da rede.

Com todas essas etapas concluídas, a LSTM foi arquitetada da maneira ilustrada na Figura 3.4.

A rede da Figura 3.4 é composta por uma camada do tipo LSTM, um número a definir de camadas ocultas e uma camada de saída com um neurônio. Em todas as camadas também utilizamos a função de ativação reLU. Implementamos o Adam para trabalhar na otimização do nosso modelo LSTM, tendo como métrica o erro médio absoluto e a taxa de aprendizado.

```
[ ]  def LSTM (hidden_layers, neurons, learning_rate):

       LSTM_model = tf.keras.models.Sequential()

       for i in range(hidden_layers):
         LSTM_layer = tf.keras.layers.LSTM(neurons, return_sequences=True)
         LSTM_model.add(LSTM_layer)

       output_layer = tf.keras.layers.Dense(1, activation='relu')
       LSTM_model.add(output_layer)

       LSTM_model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=learning_rate),
                          loss='mse', metrics=['mean_absolute_error'])

       return LSTM_model
```

Figura 3.4: Construção do LSTM através da biblioteca TensorFlow.

3.2.3 Preditor Conjunto

Adotamos uma abordagem de aprendizado conjunto para prever o tempo de trânsito do CME até chegar na Terra. Essa abordagem usa um conjunto de modelos de aprendizado de máquina cujas previsões individuais são combinadas para realizar uma previsão final (DIETTERICH, 2000). Um método *ensemble* é frequentemente mais preciso do que os modelos individuais de aprendizado de máquina que formam o método ensemble (DIETTERICH, 2000).

A estrutura de aprendizado conjunto que propomos compreende três algoritmos de aprendizado de máquina em sua estrutura, sendo eles: 1) *support vector regression* (SVR) (CORTES; VAPNIK, 1995), 2) XGBoost (CHEN, T.; GUESTRIN, 2016a) e 3) *Gaussian process* (GP) (YUAN et al., 2008). Esses algoritmos foram escolhidos, pois são utilizados em problemas de regressãocom frequência. Além disso, esses três algoritmos de regressão são comumente usados em heliofísica e pesquisa do clima espacial (LIU, C. et al., 2017; GRUET et al., 2018).

Há várias técnicas para a construção desse tipo de método. Neste trabalho, nós utilizamos a agregação *bootstrap*, originalmente proposta por Breiman (1996), que funciona da seguinte forma:

1. Selecione aleatoriamente N amostras de treinamento com reposição de um dado conjunto de treinamento com M amostras.

2. Repita a etapa (1) para gerar L subconjuntos de treinamento, $\{N_1, N_2, \ldots, N_L\}$, onde L é o número de algoritmos de aprendizado de máquina que formam o grupo do método.

Vale ressaltar que a mesma amostra pode aparecer várias vezes nos subconjuntos de treinamento.

3. Cada algoritmo do grupo é treinado individualmente por um dos subconjuntos de treinamento, $\{N_1, N_2, \ldots, N_L\}$. Sem que dois algoritmos sejam treinados pelo mesmo subconjunto de treinamento.

4. Cada algoritmo treinado faz uma previsão em um conjunto de teste pré-determinado. Assim, o método ensemble produz uma previsão final na amostra de teste através de um método de combinação, que geralmente funciona tomando a média das L previsões feitas pelos algoritmos treinados.

3.2.4 Otimização

Foram usadas duas técnicas de otimização dos modelos a fim de conseguir identificar os melhores hiperparâmetros para a MLP e LSTM: increasing window cross-validation e grid search. A inclusão dessa etapa na metodologia foi de extrema importância para evitar o sobreajuste do modelo dada a baixa quantidade de dados disponíveis.

3.2.5 Avaliação

Os resultados do trabalho foram avaliados através de uma etapa de testes utilizando 30% dos dados reais da nossa base de dados. Nessa fase, avaliamos os resultados de predição considerando as seguintes métricas de problemas de regressão: erro médio absoluto, dado pela equação Equação 3.1, erro quadrático médio, dado pela equação Equação 3.2 e raiz do erro quadrático médio, dado pela equação Equação 3.3, onde y_i representa a saída real, $\hat{y}_i$ é o valor predito e n simboliza o número de amostras no conjunto.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{3.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{3.3}$$

Capítulo 4

Resultados

Nesta seção, são apresentados os resultados experimentais obtidos neste projeto. Para essa etapa, usamos 30% dos dados da nossa base que jamais haviam sido vistos pelo modelo, na tentativa de simular o comportamento no mundo real. É importante ressaltar que a divisão da base de dados em treinamento e teste foi feita seguindo a ordem natural dos dados, ou seja, o conjunto de testes é formado pelos dados mais recentes da série temporal. Os resultados do emprego da RNA do tipo MLP são discutidos na Seção 4.1, os da aplicação da RNA do tipo LSTM na Seção 4.2 e do conjunto de preditores na Seção 4.3.

4.1 *Multi-Layer Perceptron*

A rede neural foi configurada de acordo com os valores passados ao Grid Search, sendo eles:

- 2, 4 ou 6 camadas ocultas
- 30, 50, 70 ou 90 neurônios em cada camada oculta
- 0,005 ou 0,01 como taxa de aprendizado do otimizador Adam

Além disso, ela foi configurada para receber cinco entradas e retornar uma saída. Como comentado anteriormente na Seção 3.2.5, as métricas utilizadas para avaliar os modelos foram o erro médio absoluto, erro quadrático médio e a raiz do erro quadrático médio. Dessa maneira, a Tabela 4.1 apresenta a média das métricas na etapa de validação cruzada para cada combinação possível de rede.

Tabela 4.1: Desempenho da rede neural do tipo MLP durante a etapa de Grid Search.

Camadas ocultas	neurônios	taxa de aprendizado	MAE	MSE	RMSE
2	30	0.01	58.48	3757.73	61.11
2	30	0.005	20.29	608.49	24.66
2	50	0.01	19.94	603.11	24.55
2	50	0.005	20.50	612.03	24.73
2	70	0.01	22.49	641.51	25.32
2	70	0.005	58.48	3757.73	61.11
2	90	0.01	20.17	620.11	24.90
2	90	0.005	21.89	631.06	25.12
4	30	0.01	20.46	610.78	24.71
4	30	0.005	41.77	2496.38	49.96
4	50	0.01	20.44	614.97	24.79
4	50	0.005	58.48	3757.73	61.11
4	70	0.01	21.39	620.73	24.91
4	70	0.005	21.53	621.35	24.92
4	90	0.01	20.89	617.25	24.84
4	90	0.005	21.55	628.50	25.06
6	30	0.01	19.53	601.09	24.51
6	30	0.005	58.48	3757.73	61.11
6	50	0.01	19.72	582.47	24.13
6	50	0.005	58.48	3757.73	61.11
6	70	0.01	17.74	553.09	23.51
6	70	0.005	58.48	3757.73	61.11
6	**90**	**0.01**	**16.23**	**524.35**	**22.89**
6	90	0.005	18.27	572.51	23.92

Podemos perceber que as melhores configurações de rede foram destacadas em negrito na Tabela 4.1. A que apresentou o melhor desempenho em todas as métricas possui seis camadas ocultas, noventa neurônios em cada camada e uma taxa de aprendizado de um centésimo.

4.2 *Long Short-Term Memory*

A rede neural foi configurada com os mesmos valores passados ao Grid Search para a rede neural do tipo MLP na Seção 4.1. Ela também foi configurada para receber cinco entradas

e retornar uma saída. Como comentado anteriormente na Seção 3.2.5, as métricas utilizadas para avaliar os modelos foram o erro médio absoluto, erro quadrático médio e a raiz do erro quadrático médio. Dessa maneira, a Tabela 4.2 apresenta a média das métricas na etapa de validação cruzada para cada combinação possível de rede.

Tabela 4.2: Desempenho da rede neural do tipo LSTM durante a etapa de Grid Search.

Camadas ocultas	**neurônios**	**taxa de aprendizado**	**MAE**	**MSE**	**RMSE**
2	30	0.01	16.98	401.18	20.02
2	30	0.005	16.58	389.04	19.72
2	50	0.01	15.05	404.89	20.12
2	**50**	**0.005**	14.65	**383.66**	**19.58**
2	70	0.01	15.38	422.94	20.56
2	70	0.005	15.41	430.93	20.75
2	90	0.01	15.68	438.42	20.9
2	90	0.005	58.48	3757.73	61.11
4	30	0.01	14.96	400.21	20.00
4	30	0.005	14.62	390.11	19.75
4	50	0.01	15.10	407.31	20.18
4	50	0.005	15.09	406.91	20.17
4	70	0.01	15.30	418.55	20.45
4	70	0.005	15.07	405.99	20.14
4	90	0.01	15.84	446.98	21.14
4	90	0.005	15.08	406.21	20.15
6	30	0.01	15.01	402.85	20.07
6	**30**	**0.005**	**14.49**	387.04	19.67
6	50	0.01	15.04	404.05	20.10
6	50	0.005	14.92	398.57	19.96
6	70	0.01	15.28	417.85	20.44
6	70	0.005	15.17	411.19	20.27
6	90	0.01	15.23	415.11	20.37
6	90	0.005	15.18	411.47	20.28

Podemos perceber que as melhores configurações de rede foram destacadas em negrito na tabela. A que apresentou melhor erro médio absoluto possui seis camadas ocultas, 30 neurônios em cada camada oculta e cinco milésimos como taxa de aprendizagem da rede. Já o melhor resultado considerando o erro quadrático médio e a raiz do erro quadrático médio

apresenta duas camadas ocultas, cinquenta neurônios por camada oculta e cinco milésimos como taxa de aprendizagem da rede em sua estrutura.

4.3 Conjunto de Preditores

Todos os três preditores deste conjunto foram treinados por um total de 150 épocas. O resultado do conjunto de preditores é a média dos valores preditos pelos três algoritmos de aprendizado de máquina, sendo eles o *Support Vector Regression* (SVR), o *Gaussian Process Regressor* (GPR) e o XGB *Regressor*. A Tabela 4.3 apresenta as métricas obtidas pelo conjunto de preditores já no conjunto de testes, dado que para essa etapa não separamos os dados em um conjunto de validação devido a sua implementação.

Tabela 4.3: Desempenho do Conjunto de Preditores no base de testes.

Estratégia	**MAE**	**MSE**	**RMSE**
Conjunto de Preditores	13.95	273.13	16.52

4.4 Considerações Finais

Nesse capítulo foram abordados os resultados dos experimentos envolvendo as redes neurais do tipo MLP e LSTM e o conjunto de preditores contendo o *Support Vector Regression* (SVR), o *Gaussian Process Regressor* (GPR) e o XGB *Regressor*.

Para a MLP apresentada na Seção 4.1, o melhor conjunto de métricas foi obtido utilizando a configuração com 6 camadas ocultas, 90 neurônios em cada camada oculta e 0.01 de taxa de aprendizado. Já com a LSTM da Seção 4.2, obtivemos o menor erro médio absoluto com 6 camadas, 30 neurônios e 0.005 de taxa de aprendizado, e o menor erro quadrático médio e sua raiz quadrara usando a configuração de 2 camadas, 50 neurônios em cada camada oculta e 0.005 de taxa de aprendizado. Por fim, o conjunto de preditores obteve o melhor desempenho dentre todas as abordagens ao utilizar a média dos valores preditos por cada algoritmo de aprendizado de máquina contido nele.

A Tabela 4.4 contém os resultados ao utilizarmos os modelos MLP, LSTM e o conjunto de preditores no conjunto de testes. A LSTM possui duas arquiteturas sendo representadas,

pois teve duas configurações diferentes obtendo o melhor resultado de MAE e MSE, e por conseguinte de RMSE também.

Tabela 4.4: Desempenho de todos os modelos propostos.

Estratégia	**MAE**	**MSE**	**RMSE**
MLP	17.81	602.70	24.54
LSTM I	16.52	449.40	21.19
LSTM II	16.47	447.08	21.14
Conjunto de Preditores	**13.95**	**273.13**	**16.52**

Comparando os resultados dos experimentos com a literatura, temos que o conjunto de preditores alcançou uma MAE de 13,95 horas, resultado considerado próximo à rede neural de Sudar, Vršnak e Dumbović (2015) (12 horas) e à rede de Wang et al. (2019) (12,4 horas), com diferença de 1,95 hora e 1,55 hora respectivamente. Já o desempenho da MLP e da LSTM foi levemente mais distante dos resultados dos principais modelos da literatura, o que era esperado devido ao uso de uma metodologia mais complexa nesses trabalhos.

Capítulo 5

Conclusões

Neste trabalho foram aplicados algoritmos de aprendizado de máquina para realizar a previsão do tempo de trânsito de uma CME até chegar no planeta Terra. Dentre os algoritmos usados estão o *Multi-Layer Perceptron*, o *Long Short-Term Memory* e o conjunto de preditores composto pelo *Support Vector Regression* (SVR), *Gaussian Process Regressor* (GPR) e XGB *Regressor*. A base de dados utilizada continha 726 amostras após passar pela geração de dados sintéticos, cada uma representando uma CME. A saída dos modelos correspondia ao tempo de propagação até chegar na Terra.

Com os experimentos realizados, verificamos que o conjunto de preditores foi superior às redes neurais do tipo MLP e LSTM. Além disso, notou-se que o desempenho do conjunto de preditores obteve um erro médio absoluto próximo dos modelos propostos por Sudar, Vršnak e Dumbović (2015) e Wang et al. (2019), com diferença de 1,95 hora e 1,55 hora respectivamente. Apesar disso, não foi encontrado um modelo de predição capaz de obter um erro médio absoluto de 10 horas.

Como trabalhos futuros, sugere-se a ampliação da base de dados, explorar outros parâmetros e configurações na construção das redes neurais do tipo MLP e LSTM, utilizar outros algoritmos de aprendizado de máquina no conjunto de preditores e verificar o seu desempenho, e adicionar na metodologia uma etapa de regularização para tentar lidar com o problema de poucos dados na base.

Referências bibliográficas

ABADI, M. et al. **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems**. 2015. Disponível em: <https://www.tensorflow.org>. Acesso em: 7 mai. 2023.

ADHIKARI, R.; AGRAWAL, R. K. An introductory study on time series modeling and forecasting. **arXiv preprint arXiv:1302.6613**, 2013.

AWAD, M.; KHANNA, R.; AWAD, M.; KHANNA, R. Support vector regression. **Efficient learning machines: Theories, concepts, and applications for engineers and system designers**, Springer, p. 67–80, 2015.

BAKER, D. What is space weather? **Advances in Space Research**, Elsevier BV, v. 22, n. 1, p. 7–16, jan. 1998. DOI: 10.1016/s0273-1177(97)01095-8.

BARBIERI, L. P.; MAHMOT, R. E. October-November 2003's space weather and operations lessons learned. **Space Weather**, American Geophysical Union (AGU), v. 2, n. 9, n/a–n/a, set. 2004. DOI: 10.1029/2004sw000064.

BREIMAN, L. Bagging predictors. **Machine Learning**, Springer Science e Business Media LLC, v. 24, n. 2, p. 123–140, ago. 1996. DOI: 10.1007/bf00058655.

CAMPOREALE, E. The challenge of machine learning in space weather: Nowcasting and forecasting. **Space Weather**, Wiley Online Library, v. 17, n. 8, p. 1166–1207, 2019.

CANNON, P. **Extreme space weather:impacts on engineered systems and infrastructure**. 1. ed. https://eprints.lancs.ac.uk/id/eprint/64443/: Lancaster University, 2013. ISBN 1-903496-95-0.

CARRINGTON, R. C. Description of a Singular Appearance seen in the Sun on September 1, 1859. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press (OUP), v. 20, n. 1, p. 13–15, nov. 1859. DOI: 10.1093/mnras/20.1.13.

CASTRO, M. C. F. de. **Predição não-linear de series temporais usando redes neurais RBF por decomposição em componentes principais.** 2001. Tese (Doutorado) – University of Campinas, Brazil.

CHEN, T.; GUESTRIN, C. XGBoost. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, ago. 2016. DOI: 10.1145/2939672.2939785.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. arXiv, 2016. DOI: 10.48550/ARXIV.1603.02754.

CHEN, T.; HE, T. et al. Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.

CHOLLET, F. et al. **Keras**. 2015. Disponível em: <https://github.com/fchollet/keras>. Acesso em: 7 mai. 2023.

CLETTE, F. et al. From the Wolf number to the International Sunspot Index: 25 years of SIDC. **Advances in Space Research**, Elsevier, v. 40, n. 7, p. 919–928, 2007.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer Science e Business Media LLC, v. 20, n. 3, p. 273–297, set. 1995. DOI: 10.1007/bf00994018.

DIETTERICH, T. G. Ensemble Methods in Machine Learning. In: MULTIPLE Classifier Systems. [S.l.]: Springer Berlin Heidelberg, 2000. P. 1–15. DOI: 10.1007/3-540-45014-9_1.

DRUCKER, H. et al. Support vector regression machines. **Advances in neural information processing systems**, v. 9, 1996.

DRYER, M. Interplanetary shock waves generated by solar flares. **Space Science Reviews**, Springer, v. 15, n. 4, p. 403–468, fev. 1974.

DYER, C. et al. Solar particle enhancements of single-event effect rates at aircraft altitudes. **IEEE Transactions on Nuclear Science**, Institute of Electrical e Electronics Engineers (IEEE), v. 50, n. 6, p. 2038–2045, dez. 2003. DOI: 10.1109/tns.2003.821375.

EASTWOOD, J. P. et al. The Economic Impact of Space Weather: Where Do We Stand? **Risk Analysis**, Wiley, v. 37, n. 2, p. 206–218, fev. 2017. DOI: 10.1111/risa.12765.

ECHER, E. et al. Introduction to space weather. **Advances in Space Research**, Elsevier, v. 35, n. 5, p. 855–865, 2005.

EDDY, J. A. The Maunder Minimum. **Science**, American Association for the Advancement of Science (AAAS), v. 192, n. 4245, p. 1189–1202, jun. 1976. DOI: 10.1126/science.192.4245.1189.

FENG, X. et al. A practical database method for predicting arrivals of "average" interplanetary shocks at Earth. **Journal of Geophysical Research: Space Physics**, Wiley Online Library, v. 114, A1, 2009.

FLECK, L. et al. Redes neurais artificiais: Princıpios básicos. **Revista Eletrônica Cientıfica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, 2016.

FLETCHER, L. et al. An observational overview of solar flares. **Space science reviews**, Springer, v. 159, p. 19–106, 2011.

FRY, C. et al. Forecasting solar wind structures and shock arrival times using an ensemble of models. **Journal of Geophysical Research: Space Physics**, Wiley Online Library, v. 108, A2, 2003.

FU, H. et al. Joint Geoeffectiveness and Arrival Time Prediction of CMEs by a Unified Deep Learning Framework. **Remote Sensing**, MDPI, v. 13, n. 9, p. 1738, 2021.

GAMBELLA, C.; GHADDAR, B.; NAOUM-SAWAYA, J. Optimization problems for machine learning: A survey. **European Journal of Operational Research**, Elsevier, v. 290, n. 3, p. 807–828, 2021.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GETLEY, I. L.; DULDIG, M. L.; SMART, D. F.; SHEA, M. A. Radiation dose along North American transcontinental flight paths during quiescent and disturbed geomagnetic conditions. **Space Weather**, American Geophysical Union (AGU), v. 3, n. 1, n/a–n/a, jan. 2005. DOI: 10.1029/2004sw000110.

GOLDSTON, R. J. **Introduction to plasma physics**. [S.l.]: CRC Press, 2020.

GOPALASWAMY, R.; NARAYANAN, P.; NARAYANAN, S. Cloning, overexpression, and characterization of a serine/threonine protein kinase pknI from Mycobacterium tuberculosis H37Rv. **Protein expression and purification**, Elsevier, v. 36, n. 1, p. 82–89, 2004.

GRAVES, A.; GRAVES, A. Long short-term memory. **Supervised sequence labelling with recurrent neural networks**, Springer, p. 37–45, 2012.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. **Neural networks**, Elsevier, v. 18, n. 5-6, p. 602–610, 2005.

GRUET, M. A.; CHANDORKAR, M.; SICARD, A.; CAMPOREALE, E. Multiple-Hour-Ahead Forecast of the Dst Index Using a Combination of Long Short-Term Memory Neural Network and Gaussian Process. **Space Weather**, American Geophysical Union (AGU), v. 16, n. 11, p. 1882–1896, nov. 2018. DOI: 10.1029/2018sw001898.

HAHNLOSER, R. H. et al. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. **nature**, Nature Publishing Group UK London, v. 405, n. 6789, p. 947–951, 2000.

HARA, K.; SAITO, D.; SHOUNO, H. Analysis of function of rectified linear unit used in deep learning. In: IEEE. 2015 international joint conference on neural networks (IJCNN). [S.l.: s.n.], 2015. P. 1–8.

HESS, P.; ZHANG, J. A Study of the Earth-Affecting CMEs of Solar Cycle 24. **Solar Physics**, Springer Science e Business Media LLC, v. 292, n. 6, jun. 2017. DOI: 10.1007/s11207-017-1099-y.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

HOWARD, T. **Space weather and coronal mass ejections**. [S.l.]: Springer, 2014.

HOYT, D. V.; SCHATTEN, K. H. Group sunspot numbers: A new solar activity reconstruction. **Solar physics**, Springer, v. 179, p. 189–219, 1998.

KANE, M. et al. Plasma convection in Saturn's outer magnetosphere determined from ions detected by the Cassini INCA experiment. **Geophysical Research Letters**, American Geophysical Union (AGU), v. 35, n. 4, fev. 2008. DOI: 10.1029/2007gl032342.

KELLY, M. A.; COMBERIATE, J. M.; MILLER, E. S.; PAXTON, L. J. Progress toward forecasting of space weather effects on UHF SATCOM after Operation Anaconda. **Space Weather**, American Geophysical Union (AGU), v. 12, n. 10, p. 601–611, out. 2014. DOI: 10.1002/2014sw001081.

KIM, K.-H.; MOON, Y.-J.; CHO, K.-S. Prediction of the 1-AU arrival times of CME-associated interplanetary shocks: Evaluation of an empirical interplanetary shock propagation model. **Journal of Geophysical Research: Space Physics**, Wiley Online Library, v. 112, A5, 2007.

KINGMA, D. P.; BA, J. Adam: A Method for Stochastic Optimization. **arXiv preprint arXiv:1412.6980**, arXiv, v. 9, n. 1, jan. 2014. DOI: 10.48550/ARXIV.1412.6980.

KUBAT, M. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. **The Knowledge Engineering Review**, Cambridge University Press, v. 13, n. 4, p. 409–412, 1999.

LANZEROTTI, L. et al. **Solar Physics**, Springer Science e Business Media LLC, v. 204, n. 1/2, p. 351–359, 2001. DOI: 10.1023/a:1014289410205.

LIU, C.; DENG, N.; WANG, J. T. L.; WANG, H. Predicting Solar Flares Using SDO/HMI Vector Magnetic Data Products and the Random Forest Algorithm. **The Astrophysical Journal**, American Astronomical Society, v. 843, n. 2, p. 104, jul. 2017. DOI: 10.3847/1538-4357/aa789b.

LIU, J. et al. A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CAT-PUMA. **The Astrophysical Journal**, American Astronomical Society, v. 855, n. 2, p. 109, mar. 2018. DOI: 10.3847/1538-4357/aaae69.

MCKINNEY, W. et al. pandas: a foundational Python library for data analysis and statistics. **Python for high performance and scientific computing**, Seattle, v. 14, n. 9, p. 1–9, 2011.

MICHALEK, G.; PUCHOWSKA, K.; RAMS, A. Statistical Analysis of Decimetric Radio Bursts, Flares, and Coronal Mass Ejections. **Solar Physics**, Springer, v. 257, p. 113–124, 2009.

MILLWARD, G.; BIESECKER, D.; PIZZO, V.; DE KONING, C. An operational software tool for the analysis of coronagraph images: Determining CME parameters for input into the WSA-Enlil heliospheric model. **Space Weather**, Wiley Online Library, v. 11, n. 2, p. 57–68, 2013.

MITCHELL, R.; FRANK, E. Accelerating the XGBoost algorithm using GPU computing. **PeerJ Computer Science**, PeerJ Inc., v. 3, e127, 2017.

MOZAFFARI, L.; MOZAFFARI, A.; AZAD, N. L. Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on San Francisco urban roads. **Engineering science and technology, an international journal**, Elsevier, v. 18, n. 2, p. 150–162, 2015.

MURTAGH, F. Multilayer perceptrons for classification and regression. **Neurocomputing**, Elsevier, v. 2, n. 5-6, p. 183–197, 1991.

OLIPHANT, T. E. et al. **A guide to NumPy**. [S.l.]: Trelgol Publishing USA, 2006. v. 1.

PAOURIS, E.; MAVROMICHALAKI, H. Interplanetary Coronal Mass Ejections Resulting from Earth-Directed CMEs Using SOHO and ACE Combined Data During Solar Cycle 23. **Solar Physics**, Springer Science e Business Media LLC, v. 292, n. 2, jan. 2017. DOI: 10.1007/s11207-017-1050-2.

PARKER, E. N. The Solar-Flare Phenomenon and the Theory of Reconnection and Annihiliation of Magnetic Fields. **The Astrophysical Journal Supplement Series**, American Astronomical Society, v. 8, p. 177, jul. 1963. DOI: 10.1086/190087.

PATKI, N.; WEDGE, R.; VEERAMACHANENI, K. The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). [S.l.]: IEEE, out. 2016. DOI: 10.1109/dsaa.2016.49.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011.

RAMCHOUN, H.; GHANOU, Y.; ETTAOUIL, M.; JANATI IDRISSI, M. A. Multilayer perceptron: Architecture optimization and training. International Journal of Interactive Multimedia e Artificial Intelligence ..., 2016.

RICHARDSON, I. G.; CANE, H. V. Near-Earth Interplanetary Coronal Mass Ejections During Solar Cycle 23 (1996 – 2009): Catalog and Summary of Properties. **Solar Physics**, Springer Science e Business Media LLC, v. 264, n. 1, p. 189–237, mai. 2010. DOI: 10.1007/s11207-010-9568-6.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

SALOMON, D. **Data compression: the complete reference**. [S.l.]: Springer Science & Business Media, 2004.

SCHWENN, R. Space weather: The solar perspective. **Living reviews in solar physics**, Springer, v. 3, n. 1, p. 1–72, 2006.

SEVERNY, A. Solar flares. **Annual Review of Astronomy and Astrophysics**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 2, n. 1, p. 363–400, 1964.

SHARMA, S.; SHARMA, S.; ATHAIYA, A. Activation functions in neural networks. **Towards Data Sci**, v. 6, n. 12, p. 310–316, 2017.

SHEN, C. et al. Full-halo coronal mass ejections: Arrival at the Earth. **Journal of Geophysical Research: Space Physics**, American Geophysical Union (AGU), v. 119, n. 7, p. 5107–5116, jul. 2014. DOI: 10.1002/2014ja020001.

SONG, W. An analytical model to predict the arrival time of interplanetary CMEs. **Solar Physics**, Springer, v. 261, n. 2, p. 311–320, 2010.

SUDAR, D.; VRŠNAK, B.; DUMBOVIĆ, M. Predicting coronal mass ejections transit times to Earth with neural network. **Monthly Notices of the Royal Astronomical Society**, The Royal Astronomical Society, v. 456, n. 2, p. 1542–1548, 2015.

TSURUTANI, B. T. The extreme magnetic storm of 1–2 September 1859. **Journal of Geophysical Research**, American Geophysical Union (AGU), v. 108, A7, 2003. DOI: 10.1029/2002ja009504.

VILJANEN, A. et al. Recordings of geomagnetically induced currents and a nowcasting service of the Finnish natural gas pipeline system. **Space Weather**, American Geophysical Union (AGU), v. 4, n. 10, n/a–n/a, out. 2006. DOI: 10.1029/2006sw000234.

VRŠNAK, B. et al. Propagation of interplanetary coronal mass ejections: The drag-based model. **Solar physics**, Springer, v. 285, p. 295–315, 2013.

WANG, Y.; LIU, J.; JIANG, Y.; ERDÉLYI, R. CME arrival time prediction using convolutional neural network. **The Astrophysical Journal**, IOP Publishing, v. 881, n. 1, p. 15, 2019.

WRENN, G. L.; RODGERS, D. J.; RYDEN, K. A. A solar cycle of spacecraft anomalies due to internal charging. **Annales Geophysicae**, Copernicus GmbH, v. 20, n. 7, p. 953–956, jul. 2002. DOI: 10.5194/angeo-20-953-2002.

XU, Y. et al. Artificial intelligence: A powerful paradigm for scientific research. **The Innovation**, Elsevier, v. 2, n. 4, p. 100179, 2021.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, Elsevier, v. 415, p. 295–316, 2020.

YASHIRO, S.; GOPALSWAMY, N. et al. A catalog of white light coronal mass ejections observed by the SOHO spacecraft. **Journal of Geophysical Research: Space Physics**, Wiley Online Library, v. 109, A7, 2004.

YASHIRO, S.; GOPALSWAMY, N. Statistical relationship between solar flares and coronal mass ejections. **Proceedings of the International Astronomical Union**, Cambridge University Press, v. 4, S257, p. 233–243, 2008.

YOUSSEF, M. On the relation between the CMEs and the solar flares. **NRIAG Journal of Astronomy and Geophysics**, Taylor & Francis, v. 1, n. 2, p. 172–178, 2012.

YU, Y.; ZHU, Y.; LI, S.; WAN, D. Time series outlier detection based on sliding window prediction. **Mathematical problems in Engineering**, Hindawi, v. 2014, 2014.

YUAN, J.; WANG, K.; YU, T.; FANG, M. Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression. **International Journal of Machine Tools and Manufacture**, Elsevier BV, v. 48, n. 1, p. 47–60, jan. 2008. DOI: 10.1016/j.ijmachtools.2007.07.011.

ZHANG, F.; O'DONNELL, L. J. Support vector regression. In: MACHINE learning. [S.l.]: Elsevier, 2020. P. 123–140.

ZHANG, J. et al. On the temporal relationship between coronal mass ejections and flares. **The Astrophysical Journal**, IOP Publishing, v. 559, n. 1, p. 452, 2001.

ZHAO, X.; DRYER, M. Current status of CME/shock arrival time prediction. **Space Weather**, Wiley Online Library, v. 12, n. 7, p. 448–469, 2014.