RELATÓRIO TÉCNICO VII DESAFIO EM CIÊNCIA DE DADOS: PREVISÃO DO ÍNDICE DST USANDO REDES LSTM E DADOS SOLARES

ANDREI FERREIRA INOMATA – andreiinomata36@gmail.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS - PUCGO

CAIO HENRIQUE LIMA DE ANDRADE – infinit.dev7@gmaill.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS - PUCGO

GUILHERME MAGALHÃES LIMA –
guimaglima-2004@hotmail.com PONTIFÍCIA UNIVERSIDADE
CATÓLICA DE GOIÁS - PUCGO

SOFIA SOUZA COSTA – sofiasc.pel@gmail.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS - PUCGO

VITOR MANOEL EMIDIO MACHADO – vitoremidiomachado@gmail.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS – PUCGO

MARIA JOSÉ PEREIRA DANTAS – mjpdantas@gmail.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS - PUCGO

JOSÉ ELMO DE MENEZES – jelmo.maf@gmail.com
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS - PUCGO

RESUMO: Este trabalho apresenta o desenvolvimento de um modelo preditivo para o índice Dst - Disturbance Storm Time Index (Índice de Tempo de Tempestade de Distúrbio, em tradução livre), utilizado para monitorar tempestades geomagnéticas e seus impactos no campo magnético terrestre. Como parte do VII Desafio em Ciência de Dados, foram analisados dados de vento solar, manchas solares e posições de satélites. A rede LSTM foi escolhida por sua capacidade de capturar padrões temporais complexos e relações de longo prazo nos dados sequenciais. Os dados foram preparados e organizados em séries temporais para garantir a coerência das previsões. O modelo foi avaliado por meio de métricas como RMSE (Root Mean Square Error).

PALAVRAS-CHAVES: Índice Dst; Previsão de tempestades geomagnéticas; Machine Learning; Redes LSTM; Séries Temporais; Vento Solar; Mitigação de Riscos Espaciais.

TECHNICAL REPORT VII DATA SCIENCE CHALLENGE: STD INDEX FORECASTING USING LSTM NETWORKS AND SOLAR DATA

ABSTRACT: This work presents the development of a predictive model for the Dst index - Disturbance Storm Time Index, used to monitor geomagnetic storms and their impacts on the Earth's magnetic field.As part of the VII Data Science Challenge, data on solar wind, sunspots and satellite positions were analyzed. The LSTM network was chosen for its ability to capture complex temporal patterns and long-term relationships in sequential data. The data was prepared and organized into time series to ensure the consistency of the predictions. The model was evaluated using metrics such as RMSE (Root Mean Square Error).

KEYWORDS: Dst Index; Prediction of geomagnetic storms; Machine Learning; LSTM networks; Time Series; Solar Wind; Mitigation of Space Risks.

PUC GOIÁS

1. INTRODUÇÃO

As tempestades geomagnéticas impactam significativamente o campo magnético da Terra e podem causar danos a sistemas de comunicação, redes elétricas e satélites. A previsão precisa dessas tempestades é essencial para mitigar seus impactos e garantir a segurança de infraestruturas críticas. Uma forma eficaz de monitorar essas tempestades é por meio do **índice Dst**, que mede a intensidade das variações magnéticas.

O objetivo deste projeto foi desenvolver um modelo eficiente para prever o índice Dst utilizando técnicas de aprendizado de máquina. Com isso, busca-se melhorar a capacidade de monitorar tempestades geomagnéticas e fornecer previsões úteis para mitigação de riscos associados a esses eventos.

Este trabalho foi realizado como parte do VII Desafio em Ciência de Dados e visa desenvolver um modelo preditivo utilizando aprendizado de máquina, especificamente uma **rede LSTM**. A rede foi treinada para prever o índice Dst com base em dados de vento solar, manchas solares e informações de satélites. A abordagem com LSTM permite capturar padrões temporais complexos e realizar previsões de curto prazo com alta precisão.

2. METODOLOGIA

Nesta seção, apresentamos as etapas metodológicas empregadas para o desenvolvimento do modelo de previsão do índice Dst. A metodologia seguiu um fluxo estruturado, abrangendo desde a preparação dos dados até o treinamento e avaliação do modelo.

Etapas do Desenvolvimento

1. Coleta e Preparação dos Dados

Os dados foram obtidos de diferentes fontes, incluindo:

- Vento solar: Informações sobre velocidade e direção do vento solar, obtidas de arquivos públicos.
- Manchas solares: Dados sobre a contagem de manchas solares, indicadores da atividade solar.
- Posição de satélites: Informações de trajetória dos satélites relevantes para a análise.
- Índice Dst: Indicador da intensidade das tempestades geomagnéticas, utilizado como variável alvo para o modelo.

Após a leitura dos arquivos CSV, os dados foram transformados e indexados por período e tempo. Esse processo garantiu a integridade das séries temporais e facilitou a combinação dos dados e a criação de um dataset coerente para o treinamento do modelo.

2. Desenvolvimento do Modelo

O modelo desenvolvido utiliza uma Rede Neural Long Short-Term Memory (LSTM). Essa escolha se justifica pela capacidade das LSTMs de lidar com dependências temporais, armazenando informações de eventos passados para fazer previsões mais precisas sobre futuros estados do sistema.

- Camada LSTM: Captura padrões temporais e dependências de longo prazo nos dados sequenciais.
- Camada Densa (Dense): Gera a previsão final do índice Dst.

A arquitetura possui duas camadas LSTM empilhadas e uma camada de saída densa para fornecer a previsão do índice Dst com base nos dados de entrada.

3. Treinamento do Modelo

O treinamento do modelo foi realizado utilizando:

- Função de Custo: Mean Squared Error (MSE), adequada para problemas de regressão.
- Otimização: Adam, para garantir uma convergência eficiente.
- Épocas: 17
- Batch Size: 32

Os dados foram divididos em conjuntos de treino e teste, garantindo uma avaliação justa do desempenho.

4. Avaliação do Modelo

Para medir a eficácia do modelo, utilizamos a métrica RMSE (Root Mean Square Error), que avalia a precisão da previsão.

5. Implementação e Visualização dos Resultados

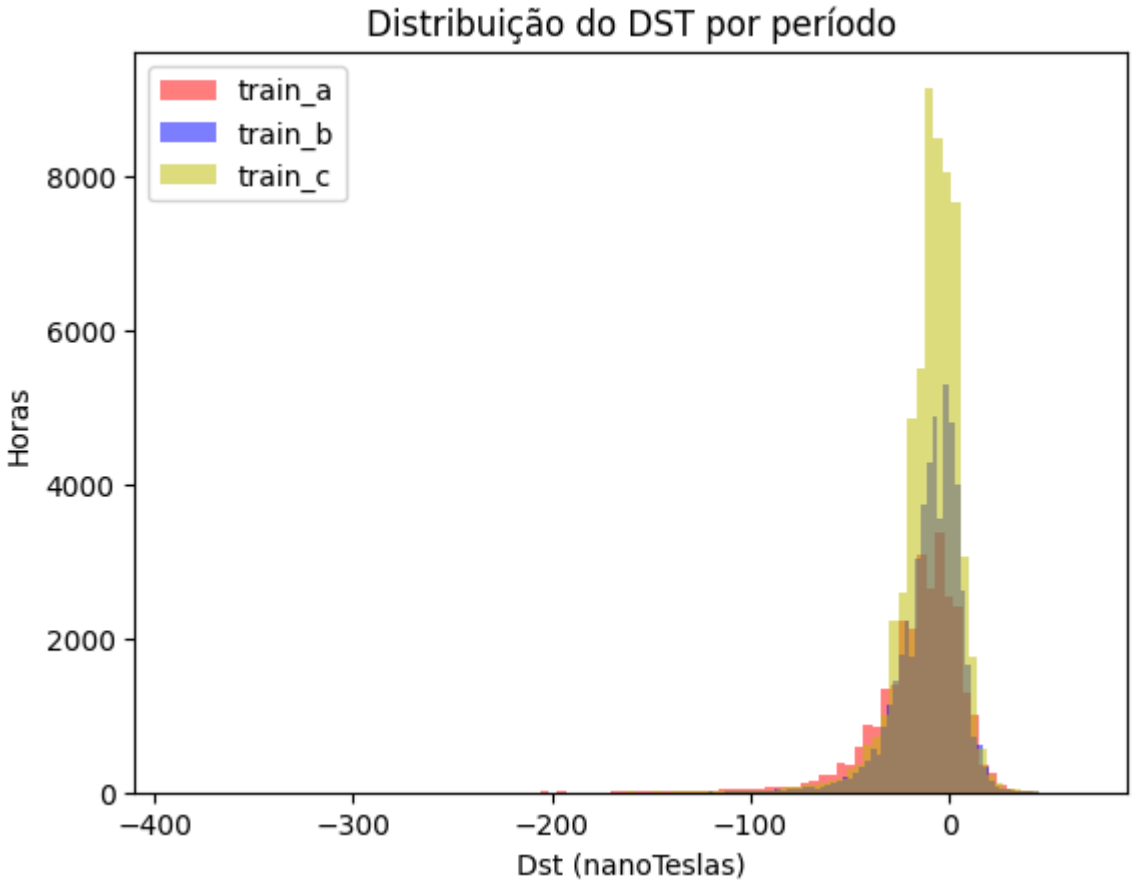
O modelo foi implementado e treinado no Google Colab devido à sua capacidade de processamento. Os resultados foram visualizados usando gráficos que mostram a comparação entre os valores previstos e reais do índice Dst, facilitando a análise de desempenho.

3. RESULTADOS E DISCUSSÕES

3.1 Histograma

O código implementa um histograma para analisar a distribuição do índice Dst ao longo de diferentes períodos. A visualização permite identificar como o índice varia em cada período, e destaca padrões e anomalias na atividade geomagnética. Cada período é representado com uma cor diferente, para facilitar a comparação visual entre eles.

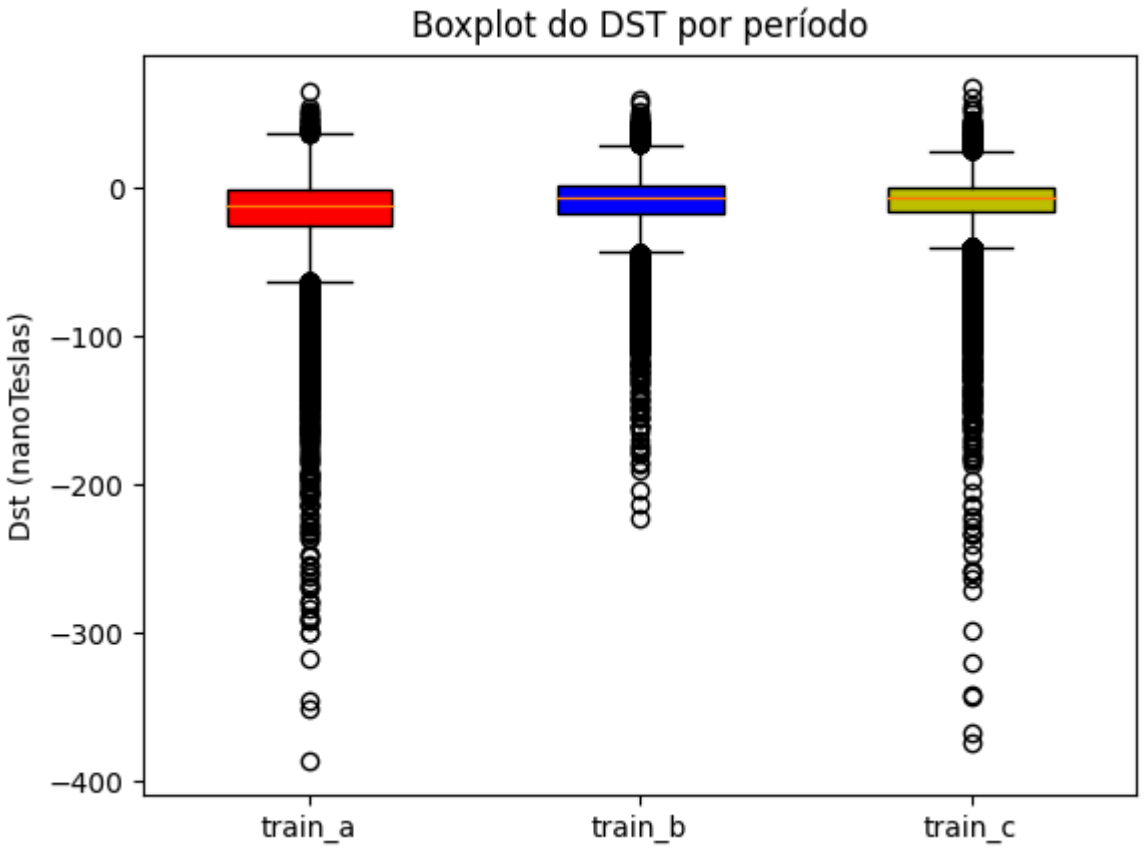
O histograma exibe o número de horas em função dos valores do índice Dst (em nanoTeslas), e oferece informações importantes sobre a intensidade e frequência das variações magnéticas em diferentes momentos. Esta análise é essencial para verificar se certos períodos apresentam maior incidência de tempestades geomagnéticas e se o comportamento do índice Dst varia significativamente ao longo do tempo.



3.2 Boxplot

Além do histograma, foi gerado um boxplot para aprofundar a análise dos dados e verificar como o índice Dst varia em diferentes períodos. O boxplot é uma ferramenta gráfica que resume a distribuição dos valores em termos de mediana, quartis e dispersão, além de destacar outliers que podem representar eventos geomagnéticos extremos. No código implementado, cada período foi representado por um boxplot separado, com uma cor distinta para facilitar a comparação visual.

O eixo vertical exibe os valores do índice Dst em nanoTeslas, enquanto o eixo horizontal apresenta os rótulos dos períodos analisados. Essa visualização é essencial para identificar quais períodos apresentam maior variabilidade no índice Dst e como essa variação se comporta ao longo do tempo. A presença de outliers pode indicar eventos significativos, como tempestades geomagnéticas intensas, que são de grande interesse no contexto da previsão e mitigação dos impactos do clima espacial.

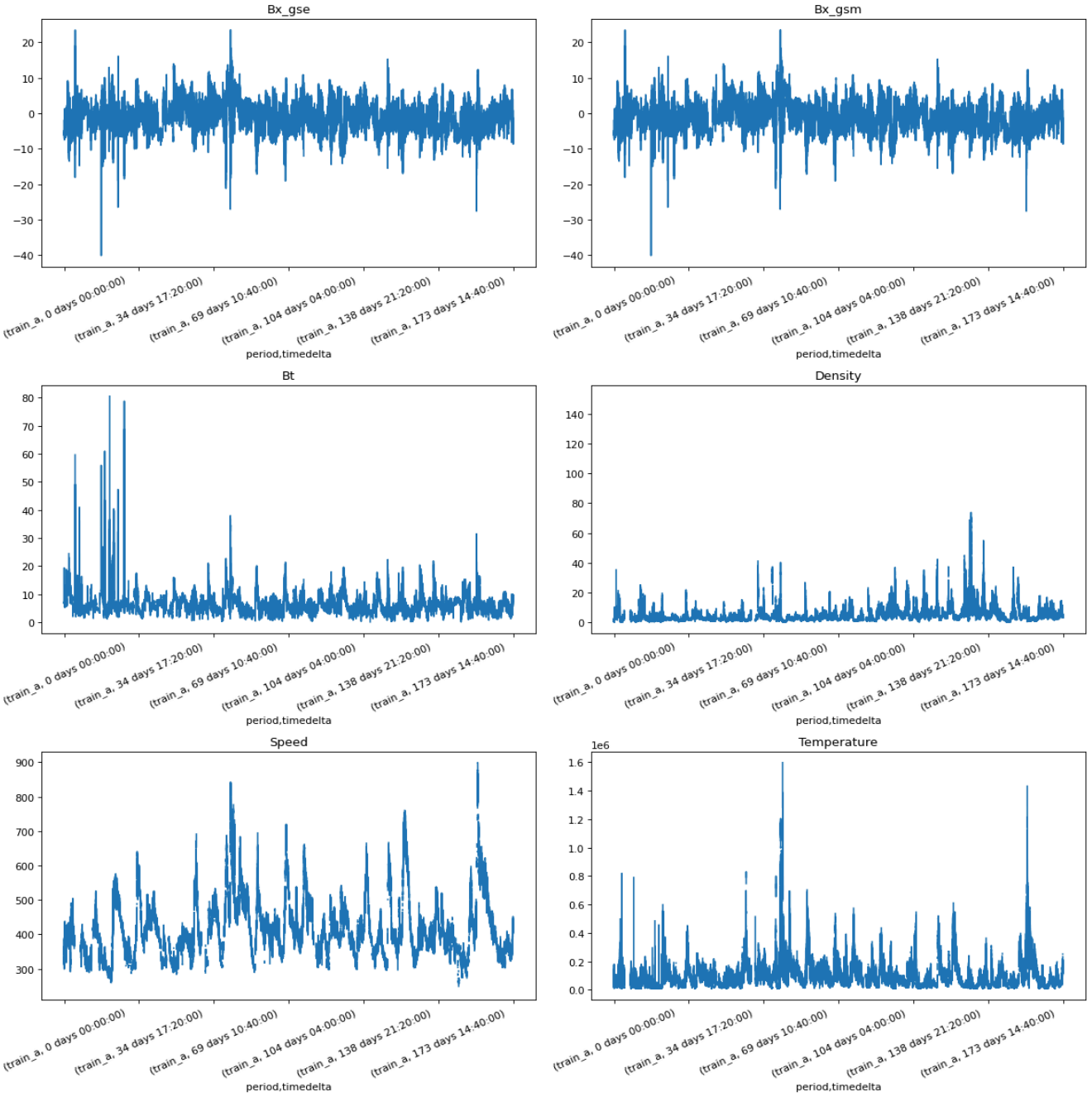


3.3 Variação das variáveis

Foi implementada uma função que gera gráficos para visualizar dados brutos relevantes à previsão do índice Dst. Essa visualização é essencial para explorar os dados de entrada, identificar padrões iniciais e validar a qualidade dos dados utilizados no treinamento do modelo LSTM. A função foca em atributos críticos relacionados ao vento solar, como densidade, velocidade e temperatura, que são conhecidos por afetar o campo magnético terrestre e, consequentemente, o índice Dst.

A função chamada show_raw_visualization cria múltiplos gráficos, permitindo uma análise rápida e detalhada dos parâmetros do vento solar armazenados na variável cols_to_plot. Entre as colunas analisadas estão variáveis como "bx_gse" e "bx_gsm" (componentes do campo magnético interplanetário), "density" (densidade do vento solar), "speed" (velocidade do vento solar), e "temperature" (temperatura das partículas do vento solar). Cada um desses atributos é exibido em um gráfico separado, facilitando a análise comparativa entre eles.

Ao exibir os dados brutos, essa função permite identificar possíveis anomalias, padrões sazonais ou mudanças bruscas nos valores que podem indicar tempestades geomagnéticas iminentes. A análise prévia dos dados por meio de gráficos é essencial para garantir que o modelo LSTM seja treinado com informações consistentes e que padrões críticos não sejam perdidos. Essa abordagem também auxilia na compreensão da relação entre o comportamento do vento solar e as variações no índice Dst, fornecendo uma base visual sólida para as etapas subsequentes de modelagem e previsão.

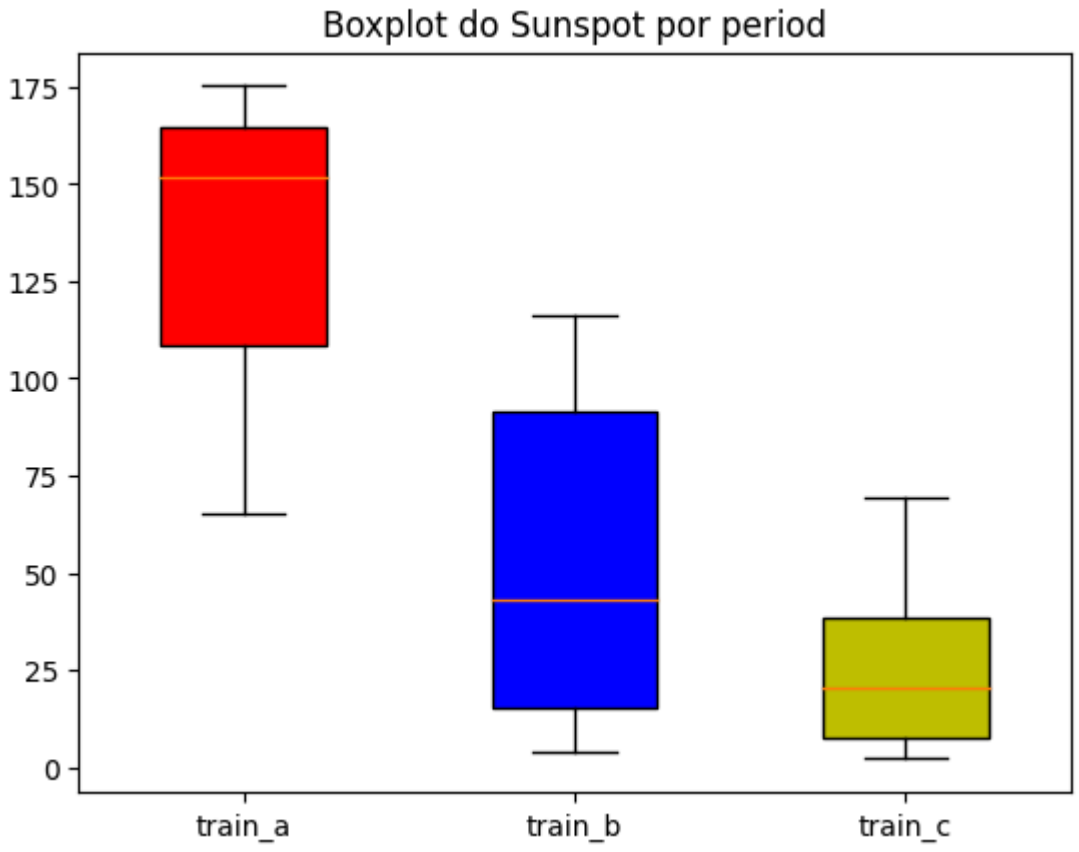


3.4 Boxplot manchas solares

Geramos um boxplot para a variável "sunspot" (manchas solares) para analisar a distribuição dessa variável ao longo de diferentes períodos. As manchas solares são indicadoras importantes da atividade solar, e suas variações ao longo do tempo estão diretamente relacionadas ao comportamento do clima espacial e às tempestades geomagnéticas, que impactam o índice Dst.

Neste boxplot, os dados de manchas solares são agrupados por período, com cada grupo representado por uma caixa de cor distinta. Essa representação gráfica facilita a comparação entre os períodos e permite observar tendências centrais (como a mediana) e a dispersão dos dados (representada pelos quartis). Além disso, a presença de outliers pode indicar eventos de atividade solar intensa, o que é relevante para antecipar tempestades geomagnéticas.

O eixo horizontal contém os rótulos dos períodos analisados (train_a, train_b e train_c), enquanto o eixo vertical mostra a contagem de manchas solares em cada período. Cada boxplot resume as variações da atividade solar durante o período correspondente, fornecendo uma visão clara da distribuição dos dados e identificando possíveis anomalias ou variações extremas.



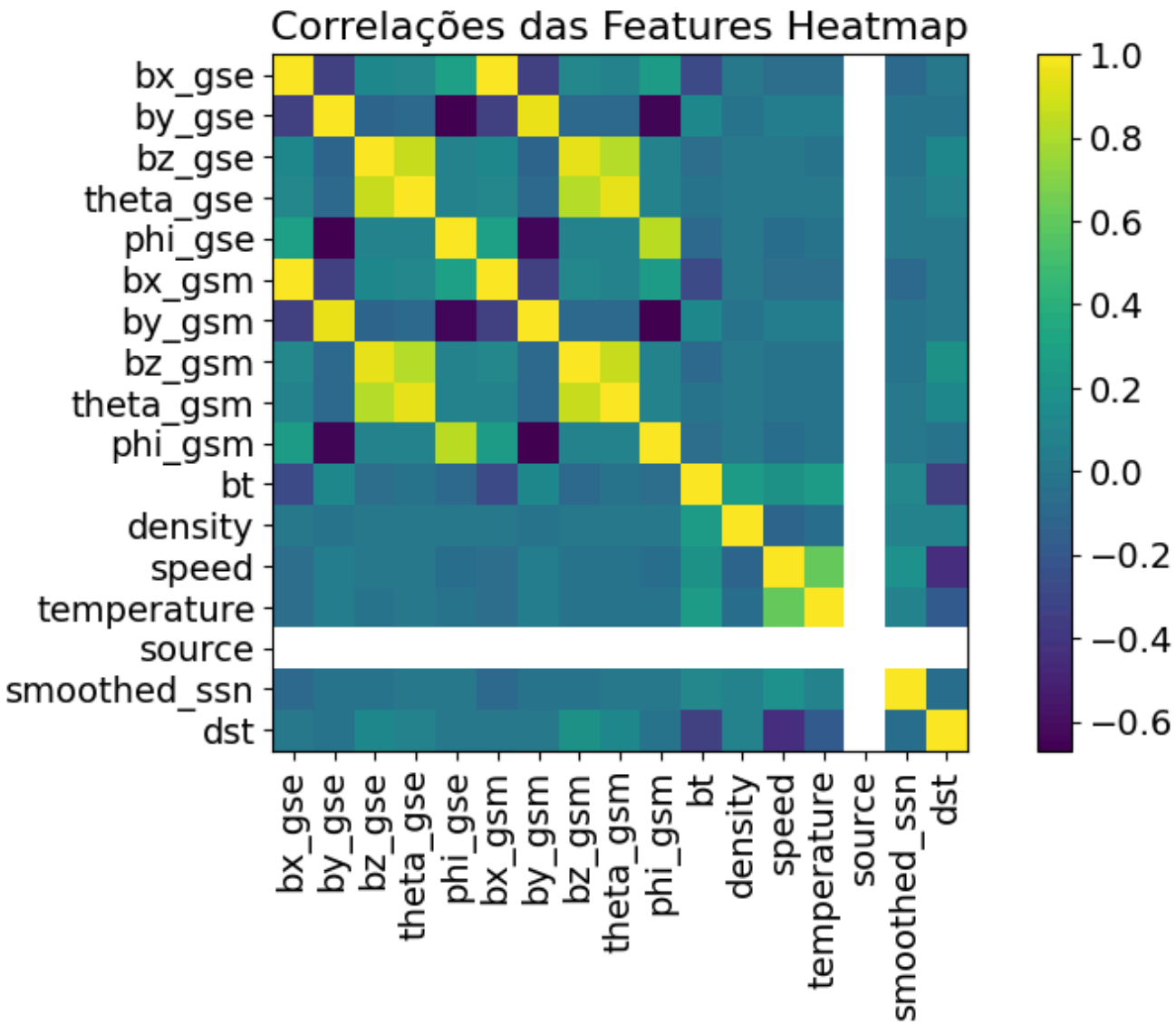
3.5 Correlações

Esse trecho de código gera um heatmap de correlações para visualizar a relação entre as variáveis dos dados de vento solar, manchas solares e o índice Dst. A análise de correlação é fundamental no contexto do artigo, pois permite identificar relações lineares entre as variáveis de entrada e o índice Dst, facilitando a seleção de atributos relevantes para a construção do modelo preditivo com LSTM.

Primeiro, os dados de vento solar, manchas solares e índice Dst são combinados em um único DataFrame por meio do método `join()`. Como pode haver valores ausentes nos dados, o método `fillna(method="ffill")` é utilizado para preencher esses valores com o último valor válido, o que garante a consistência dos dados para a análise.

O heatmap gerado exibe a matriz de correlação entre as variáveis. Cada célula do gráfico representa o coeficiente de correlação entre duas variáveis. Valores próximos de 1 indicam uma forte correlação positiva, enquanto valores próximos de -1 indicam uma forte correlação negativa. Valores próximos de 0 indicam que não há correlação linear significativa entre as variáveis.

O eixo horizontal e o eixo vertical contêm os nomes das variáveis analisadas, como componentes do vento solar, contagem de manchas solares e valores do índice Dst. O colorbar à direita fornece uma escala visual para interpretar a força das correlações. Essa análise foi importante para identificar quais variáveis têm maior impacto no comportamento do índice Dst.



3.6 Divisão dos dados - treino e teste

Criamos um gráfico de barras horizontais para visualizar a divisão dos dados entre os conjuntos de treinamento, validação e teste ao longo de diferentes períodos. Essa visualização permite verificar como os dados estão distribuídos entre as fases de treino, validação e teste, o que garante que cada período seja representado adequadamente em todas as etapas do modelo preditivo.

As variáveis `train_cnts`, `val_cnts` e `test_cnts` armazenam a contagem de amostras para cada período nos conjuntos de treinamento, validação e teste, respectivamente. A divisão dos

dados é feita com base no agrupamento por "period". Cada barra horizontal representa um período (train_a, train_b e train_c), com diferentes partes da barra correspondendo à quantidade de amostras em cada conjunto (treinamento, validação e teste).

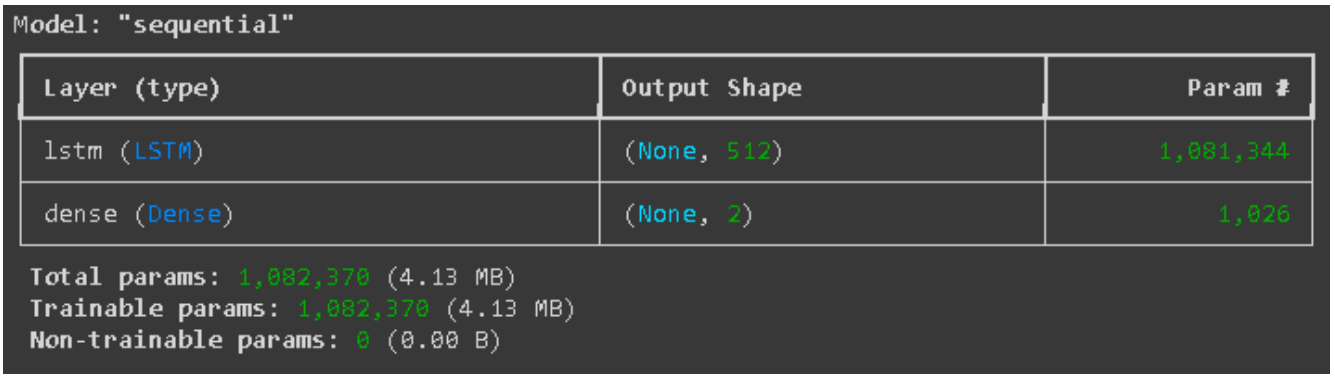
A legenda indica as diferentes fases do modelo: "Train", "Validation" e "Test". O eixo vertical contém os nomes dos períodos (train_a, train_b e train_c), enquanto o eixo horizontal mostra a quantidade de amostras em cada fase, medida em timesteps horários.

Essa visualização ajuda a garantir que cada período tenha uma distribuição balanceada entre treino, validação e teste. Isso é fundamental para evitar overfitting e garantir que o modelo LSTM generalize bem para novos dados, o que melhora a precisão e a robustez das previsões do índice Dst.

3.7 Parâmetros do modelo

Foi desenvolvido um modelo LSTM utilizando a biblioteca Keras para capturar padrões temporais nos dados e prever o comportamento do índice Dst. A arquitetura do modelo consiste em uma camada LSTM com 1024 neurônios, configurada para processar dados sequenciais e identificar dependências temporais complexas. Além disso, foi aplicada uma camada densa para gerar as previsões finais. A rede foi configurada com dropout, uma técnica que desliga aleatoriamente 30% dos neurônios durante o treinamento, visando melhorar a generalização e evitar overfitting.

O treinamento do modelo foi realizado com o otimizador RMSprop, uma variante do gradiente descendente adaptada para séries temporais, utilizando uma taxa de aprendizado de 0,001. A função de custo escolhida foi o Mean Squared Error (MSE), adequada para medir a precisão das previsões. A estrutura do modelo foi organizada de forma a garantir que ele consiga generalizar bem para novos dados e fazer previsões precisas do índice Dst, antecipando tempestades geomagnéticas e fornecendo informações úteis para a mitigação de seus impactos.


Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 512)	1,081,344
dense (Dense)	(None, 2)	1,026

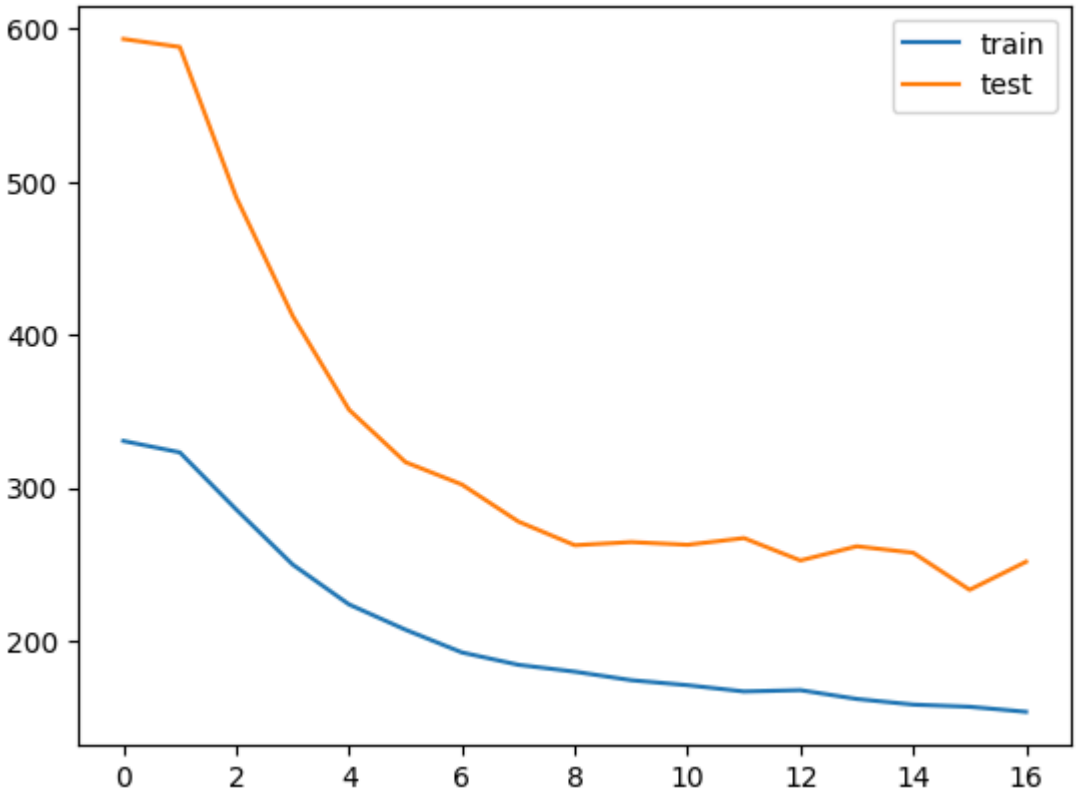
Total params: 1,082,370 (4.13 MB)
Trainable params: 1,082,370 (4.13 MB)
Non-trainable params: 0 (0.00 B)

3.8 Treinando o Modelo

No processo de treinamento do modelo LSTM, foi utilizado o método `fit()`, que ajusta os parâmetros da rede neural com base nos dados de treinamento fornecidos. Nesse processo, o modelo é alimentado com o conjunto de dados de treino (`train_ds`) e configurado para executar o treinamento por um número específico de épocas, definido em 17, conforme o valor de `n_epochs`. O treinamento foi realizado com batches, que são pequenos subconjuntos de dados processados por vez, conforme o tamanho definido na configuração. Essa abordagem permite otimizar o processo de aprendizado e melhorar a eficiência computacional.

Para garantir uma avaliação contínua do desempenho do modelo e evitar overfitting, foi utilizado um conjunto de validação (`val_ds`). Durante cada época, o modelo ajusta seus pesos com base nos dados de treinamento e, ao final de cada ciclo, avalia seu desempenho no conjunto de validação. Esse processo garante que o modelo não apenas memorize os dados de treino, mas também generalize bem para dados não vistos.

A opção `shuffle=False` foi utilizada para preservar a ordem sequencial dos dados durante o treinamento, essencial para modelos que dependem de séries temporais, como a LSTM. Com essa configuração, o treinamento segue uma ordem temporal coerente, respeitando a sequência cronológica dos eventos, o que é fundamental para melhorar a precisão das previsões do índice Dst.




```
558/558 ━━━━━━━━━━━━━━━━━━━━ 4s 7ms/step - loss: 374.7839
Test RMSE: 14.87
```

Após a obtenção do gráfico de Treino e Teste juntamente com o RMSE, o modelo já está pronto para ser aplicado na previsão de dados futuros.

4. CONCLUSÃO

Este trabalho demonstra a viabilidade do uso de redes LSTM para previsão do índice Dst, e fornece uma ferramenta útil para o monitoramento de tempestades geomagnéticas. As técnicas desenvolvidas têm potencial para aplicações futuras na mitigação de riscos espaciais e na previsão climática. Em trabalhos futuros, exploraremos arquiteturas mais sofisticadas, como redes LSTM empilhadas e modelos híbridos com Transformer.

4. CÓDIGOS UTILIZADOS

Os códigos utilizados nesta pesquisa podem ser acessados em:

https://github.com/CaioHenri99/Predicao-Eventos-Extremos

5. REFERÊNCIAS

CHUNG, C. How to Predict Disturbances in the Geomagentic Field with LSTMs - Benchmark. Disponível em: <https://drivendata.co/blog/model-geomagnetic-field-benchmark>. Acesso em: 16 out. 2024.

WITH, F. Time Series Forecasting With RNN(LSTM)| Complete Python Tutorial|. Disponível em: <https://youtu.be/S8tpSG6Q2H0?si=6_IGtJlePXBzUg_7>. Acesso em: 15 out. 2024.

BROWNLEE, J. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. Disponível em: <https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/>.

PYTORCH TUTORIAL - RNN & LSTM & GRU - RECURRENT NEURAL NETS. PyTorch Tutorial - RNN & LSTM & GRU - Recurrent Neural Nets. Disponível em: <https://youtu.be/0_PgWWmauHk?si=XpMfgmXJaY6OPMEh>. Acesso em: 15 out. 2024.


PUC
GOIÁS