UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DISCIPLINA: SCC0277 - COMPETIÇÕES DE CIÊNCIAS DE DADOS
DOCENTE: FERNANDO PEREIRA DOS SANTOS

CAMILA SAYAKA HIURA - N°USP 11218323

RELATÓRIO - DESAFIO DE SÉRIES TEMPORAIS

SÃO CARLOS - SP
2023.6

SUMÁRIO

1.1 INTRODUÇÃO

Uma análise de séries temporais tem como objetivo entender e modelar o padrão temporal subjacente nos dados, identificando tendências, padrões sazonais e quaisquer outros componentes, permitindo prever valores futuros ou analisar o impacto de certos eventos passados. Neste relatório, é apresentado um trabalho de séries temporais utilizando a linguagem de programação Python, realizado com o objetivo de prever a temperatura média da cidade de Delhi, na Índia. Para isso, foram utilizados os conjuntos de dados DailyDelhiClimateTrain e DailyDelhiClimateTest disponibilizados pelo Kaggle, plataforma de Aprendizado de Máquina e Ciência de Dados que oferece diversos conjuntos de dados públicos para desafios e análises de dados. Assim, será apresentado as etapas do processo e os resultados obtidos.

1.2 OBJETIVO

O trabalho tem como objetivo prever a temperatura média da cidade de Delhi, apresentando uma análise detalhada do conjunto de dados e descrição das transformações de dados realizadas.

1.3 DESCRIÇÃO DOS CONJUNTOS DE DADOS

O conjunto de dados DailyDelhiClimateTrain possui 1462 registros e o conjunto DailyDelhiClimateTest apresenta 114 registros, ambos com 5 colunas :

- **date**: Data do registro
- **meantemp**: Temperatura média calculada a partir de múltiplos intervalos de 3 horas em um dia
- **humidity**: Valor de umidade para o dia (unidades são gramas de vapor de água por metro cúbico de volume de ar)
- **wind_speed**: Velocidade do vento medida em km/h
- **meanpressure**: Leitura de pressão do clima (medida em atm)

Link do repositório

1.4 ANÁLISE EXPLORATÓRIA DOS DADOS

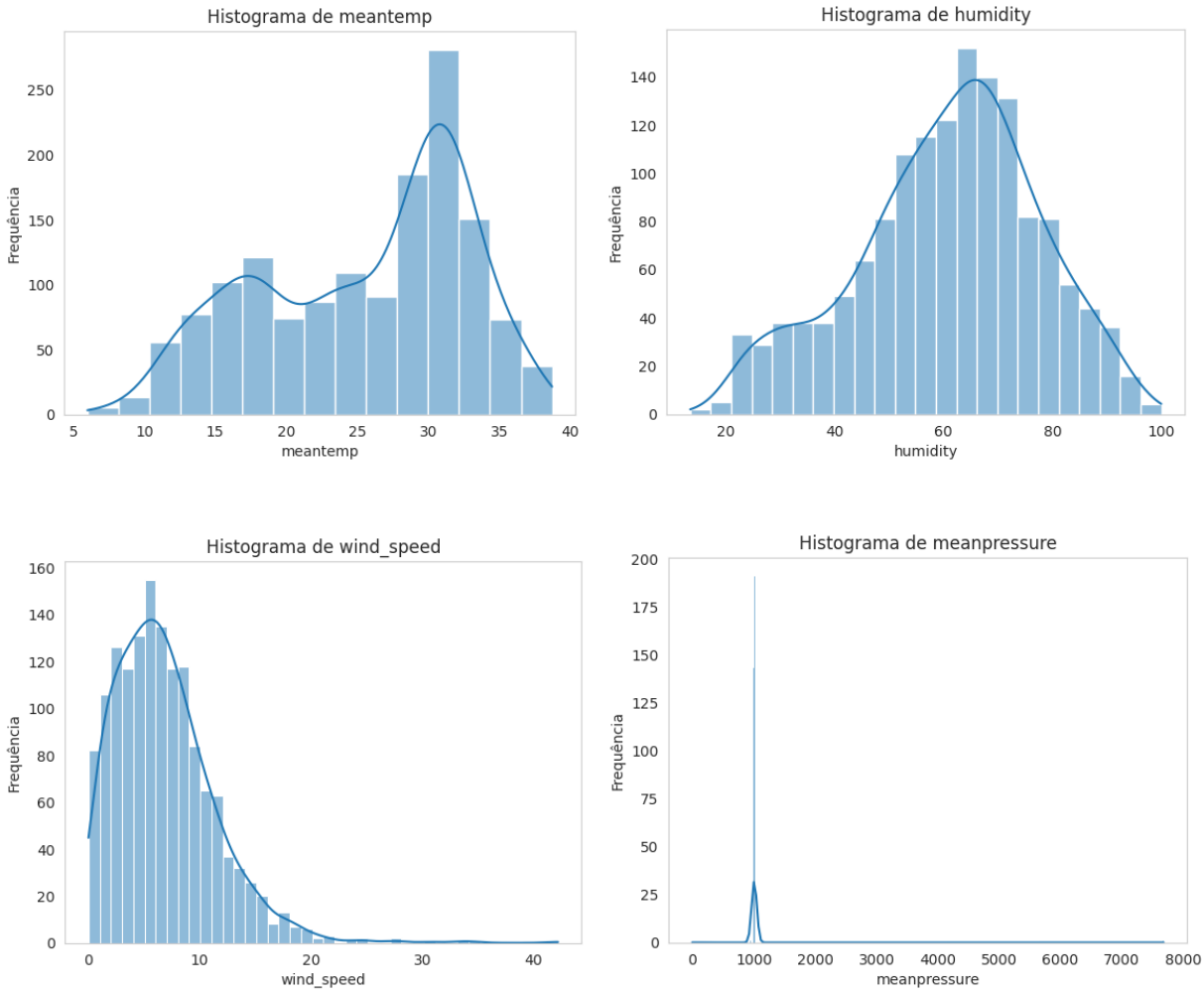
Para entender esses dados, foi realizada uma manipulação dos dados por meio do pacote Pandas em Python. Usando o comando ".describe()" nos dados, algumas estatísticas descritivas foram geradas, incluindo contagem, média, desvio-padrão, valores mínimos e máximos e quartis. A Figura 1 ilustra a saída da função ".describe()" apresentando essas estatísticas descritivas.

Figura 1 – Estatísticas descritivas das colunas numéricas

	meantemp	humidity	wind_speed	meanpressure
count	1462.000000	1462.000000	1462.000000	1462.000000
mean	25.495521	60.771702	6.802209	1011.104548
std	7.348103	16.769652	4.561602	180.231668
min	6.000000	13.428571	0.000000	-3.041667
25%	18.857143	50.375000	3.475000	1001.580357
50%	27.714286	62.625000	6.221667	1008.563492
75%	31.305804	72.218750	9.238235	1014.944901
max	38.714286	100.000000	42.220000	7679.333333

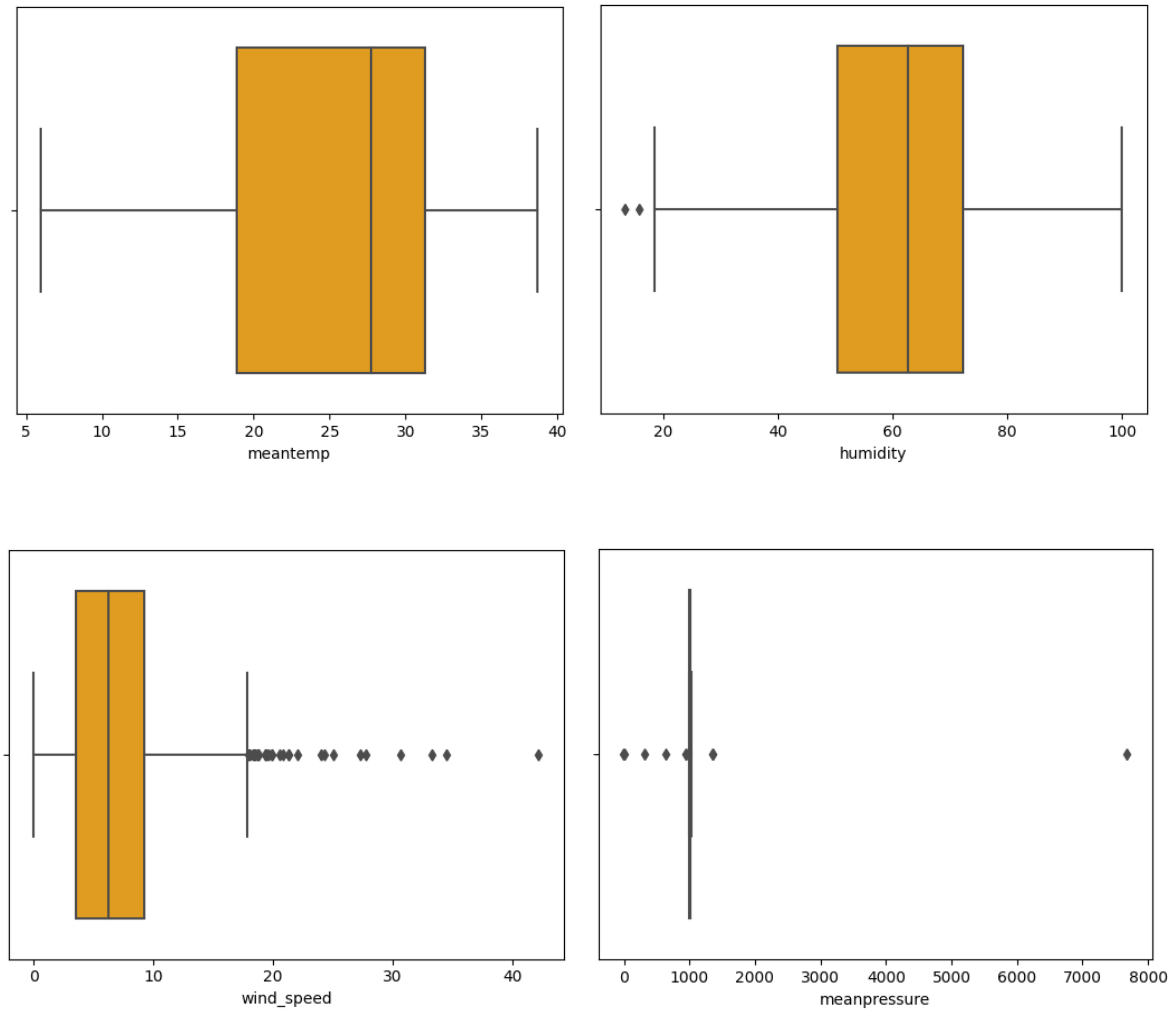
A Figura 2 apresenta os histogramas das variáveis numéricas, mostrando a distribuição desses dados. Nenhuma apresenta simetria.

Figura 2 – Histogramas das Variáveis Numéricas



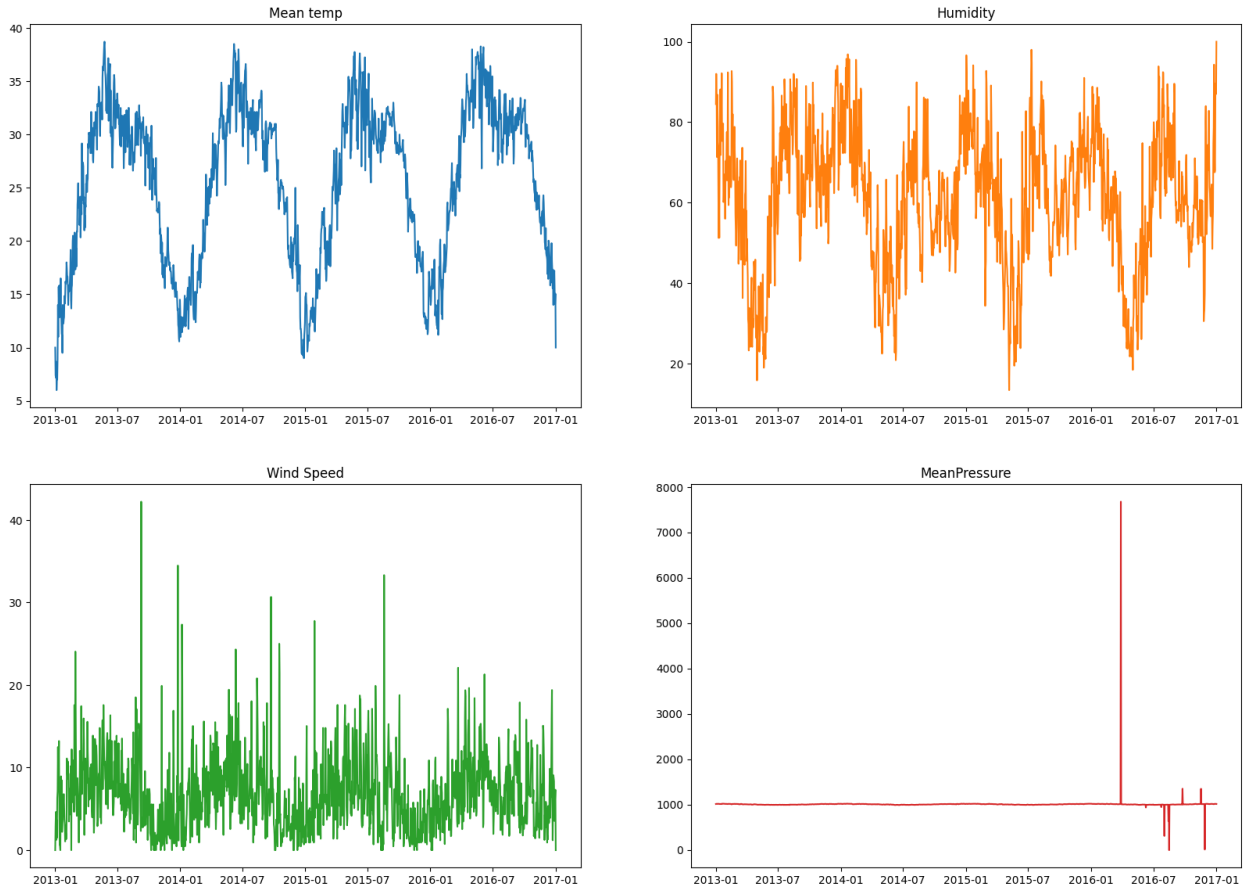
Na Figura 3, tem-se os boxplots de cada variável numérica. É perceptível a presença de outliers em humidity, wind_speed e meanpressure.

Figura 3 - Boxplots das Variáveis Numéricas



A Figura 4 mostra como os atributos se comportam ao longo do tempo. Percebe-se que a variável meanpressure apresenta outliers significativos. As variáveis meantemp e humidity aparentam ter uma certa sazonalidade e ciclo.

Figura 4 - Gráficos das Variáveis em Função do Tempo



2 DESCRIÇÃO DAS TÉCNICAS UTILIZADAS

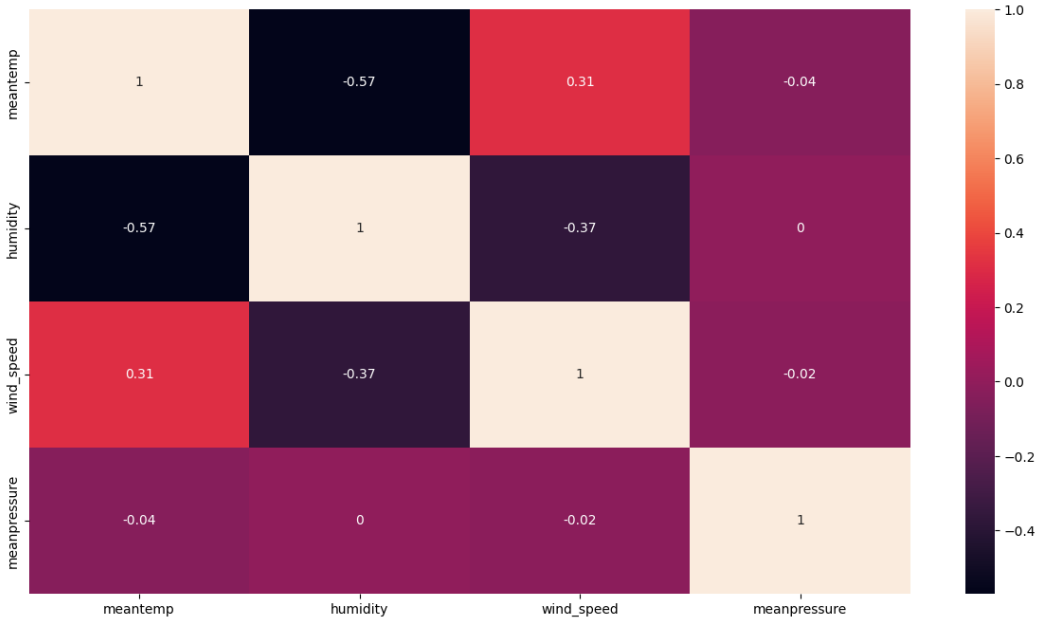
2.1 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados é uma etapa importante na análise de dados, que envolve a limpeza, transformação e organização dos dados brutos. Durante o tratamento, os dados não apresentaram linhas duplicadas, nem valores ausentes.

Para entender as relações entre as variáveis e ajudar a identificar padrões e tendências nos dados, é utilizado a matriz de correlação, uma tabela que mostra as correlações (relacionamentos estatísticos) entre um conjunto de variáveis aleatórias. É uma forma de medir a força e a direção da relação linear entre pares de variáveis em um conjunto de dados. A matriz de correlação é uma matriz quadrada onde cada célula mostra o coeficiente de correlação entre duas variáveis. O coeficiente de correlação pode variar de -1 a +1, onde -1 indica uma correlação perfeitamente negativa (ou inversa), 0 indica nenhuma correlação e +1 indica uma correlação perfeitamente positiva.

Na figura a seguir, a matriz não parece apresentar variáveis correlacionadas.

Figura 5 - Matriz de correlação



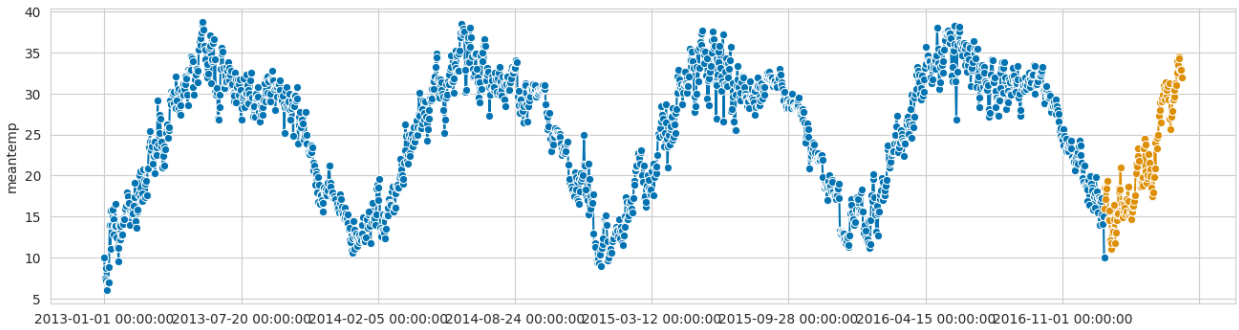
2.2 ALGORITMOS DE SÉRIES TEMPORAIS

O conjunto de dados DailyDelhiClimateTrain foi utilizado como conjunto treino e o conjunto DailyDelhiClimateTrest foi utilizado como conjunto teste.

As métricas utilizadas foram:

- Mean Absolute Error (MAE): calcula a média das diferenças absolutas entre as previsões do modelo e os valores reais.
- Mean Absolute Percentage Error (MAPE): calcula a média das porcentagens absolutas de erros em relação aos valores reais.
- Root Mean Squared Error (RMSE): raiz quadrada do MSE e representa uma medida de erro médio em unidades originais

Figura 6: Série Temporal de meantemp



2.2.1 NAIVE FORECASTER

Abordagem simples que utiliza apenas o valor mais recente da série como estimativa para o próximo período. É considerado "ingênuo" porque não leva em conta nenhum padrão, tendência ou sazonalidade nos dados.

Métricas	Last	Mean	Drift
MAE	11.71	6.61	11.71
MAPE	0.49	0.36	0.49
RMSE	13.31	7.37	13.31

2.2.2 EXPONENTIAL SMOOTHING

Tem como objetivo capturar e prever os padrões ou tendências subjacentes nos dados, atribuindo pesos às observações passadas. A ideia básica é dar mais peso às observações recentes, enquanto diminui gradualmente o peso das observações mais antigas.

MAE	10.44
MAPE	0.43
RMSE	12.19

2.2.3 AUTO ETS

Método de previsão que combina os componentes de erro, tendência e sazonalidade em uma abordagem exponencial. É uma extensão do modelo de suavização exponencial, que busca capturar as características da série temporal e fazer previsões com base nos padrões históricos observados

MAE	10.63
MAPE	0.44
RMSE	12.37

2.2.4 AUTOARIMA

Método de previsão de séries temporais que automatiza a seleção dos parâmetros do modelo ARIMA. O ARIMA é um modelo estatístico que combina componentes de autocorrelação (AR), diferenciação (I) e média móvel (MA) para modelar padrões e tendências em séries temporais.

MAE	9.41
MAPE	0.38
RMSE	11.3

2.2.5 PROPHET

Biblioteca de previsão de Séries Temporais desenvolvida pelo Facebook, que simplifica o processo de modelagem e previsão de dados temporais.

MAE	2.24
MAPE	0.12
RMSE	2.77

2.2.6 LSTM LAYER

Tipo de camada recorrente usada em redes neurais para processar sequências de dados, como séries temporais. Ela é projetada para lidar com o problema de desvanecimento do gradiente, permitindo que as redes neurais capturem dependências de longo prazo em dados sequenciais.

MAE	0.03
MAPE	0.09
RMSE	0.05

3 CONCLUSÃO

Em resumo, este estudo teve como objetivo prever a temperatura média da cidade de Delhi, na Índia. Foram aplicados diversos algoritmos, como Naive Forecaster, Exponential Smoothing, AutoETS, AutoArima, Prophet e LSTM Layer. As métricas utilizadas foram MAE, MAPE e RMSE.

Percebe-se que Naive Forecaster apresentou um baixo desempenho, enquanto Prophet e LSTM Layer obtiveram melhor resultado.

4 APÊNDICE

4.1 DÉBORA BUZON DA SILVA

O artigo propõe o Prophet, um modelo de previsão desenvolvido pela Facebook, projetado para fornecer previsões precisas e eficientes para uma ampla variedade de dados de séries temporais, podendo ser decomposto em 3 componentes principais: tendência, sazonalidade e feriados. Tem como objetivos: facilitar a produção de previsões em escala, lidar com um grande número de previsões e de séries temporais diferentes e avaliar as previsões geradas.

4.2 FELIPE CADAVEZ OLIVEIRA

O artigo propõe uma maneira de analisar espectrogramas de áudio. A solução proposta é o uso do transformador separável, que consiste em dividir a imagem em tokens de altura igual a intervalos constantes da frequência e largura igual a intervalos constantes de tempo. Ou seja, o espectrograma é dividido em tokens, que são separados em colunas de mesmo tempo e assim, é realizado um agrupamento médio dos tokens, em que são separados em intervalos de mesma frequência e incorporados mais uma vez com dados de posição resultando num dataset de tokens, onde cada dataframe é um intervalo de tempo e frequência.

4.3 GUSTAVO BARTHOLOMEU TRAD SOUZA

O artigo propõe o uso de métodos baseados em Transformers, apresentando um estudo relacionado a epidemias de influenza, destacando o uso de mecanismos de atenção.