Análise de Dados para Previsão Climática: Melhorando Modelos Preditivos com Lag-Features

1. Trabalho Inicial

Nesse trabalho, eu iniciei a análise e a previsão de dados climáticos utilizando um conjunto de dados da cidade de Presidente Prudente. As principais etapas que eu realizei foram:

Carregamento e Preparação dos Dados

- Os dados foram carregados a partir de um arquivo CSV (dados_A707_D_2014-01-01_2025-05-01.csv), removendo manualmente as 10 primeiras linhas de metadados.
- Nomeei as colunas para facilitar a manipulação, ficando com: data, precipitacao_total, pressao_atm_media, temp_orvalho_media, temp_maxima, temp_media, temp_minima, umidade_relativa_media, umidade_relativa_minima, umidade_relativa_maxima e vento_vel_media.
- Depois a coluna data foi convertida para o formato datetime.
- Tratei os valores ausentes das colunas numéricas por interpolação temporal, utilizando o método "time" e tendo a coluna data como índice.

Análise Exploratória e Visualização

- Gerei gráficos de séries temporais para visualizar a evolução de temp_media e umidade_relativa_media ao longo do tempo.
- Construí um mapa de calor (heatmap) para analisar as correlações entre as variáveis meteorológicas, após remover a coluna data.

Modelagem Preditiva (Baseline)

- Treinei um modelo de Regressão Linear para prever a temp_media.
- Para isso, escolhi as seguintes variáveis preditoras (features): temp_minima, temp_maxima, umidade_relativa_media e pressao_atm_media.
- Dividi os dados em conjuntos de treino (80%) e teste (20%).
- Em seguida, eu treinei o modelo com os dados de treino.

Avaliação do Modelo

- Avaliei o desempenho do modelo utilizando as métricas RMSE (Raiz do Erro Quadrático Médio) e R^2 (Coeficiente de Determinação).
- Os resultados mostraram um RMSE de aproximadamente 0.60 (valor aproximado, pois o texto original não tinha o número exato) e um R^2 de 0.96.
- Também gerei um gráfico de dispersão comparando os valores reais com os valores previstos, e percebi que a maioria dos pontos estava próxima da linha de identidade (y = x), indicando um bom ajuste e boa generalização, com poucos outliers visíveis.

2. Comparação de Desempenho: Modelos com Lag Features

Após o desenvolvimento do modelo baseline, comecei a investigar o impacto da inclusão de **lag features** (variáveis defasadas temporalmente) sobre a performance da previsão de temp_media. Implementei defasagens de 1, 2, 3 e 7 dias para as variáveis temp_media, temp_minima, temp_maxima, umidade_relativa_media e pressao_atm_media.

Com essas novas features, eu treinei e avaliei quatro modelos diferentes:

Modelo	**RMSE**	**R^2**
Regressão Linear (com Lags)	0.5278	0.9748
Random Forest Regressor	0.5703	0.9705
Gradient Boosting Regressor	0.5677	0.9708
Support Vector Regressor (SVR)	0.5992	0.9675

Análise Comparativa

- **Melhoria com Lag Features (Regressão Linear):**

Comparando com o modelo baseline (R^2 de 0.96), percebi uma melhora significativa. O modelo com lags teve R^2 de 0.9748 e RMSE de 0.5278, indicando que a inclusão de informações temporais passadas ajudou a capturar melhor as dependências temporais dos dados.

- **Outros Modelos com Lag Features:**
  - O **Gradient Boosting Regressor** teve um ótimo desempenho, com R^2 de 0.9708 e RMSE de 0.5677.
  - O **Random Forest Regressor** ficou muito próximo, com R^2 de 0.9705 e RMSE de 0.5703.
  - O **Support Vector Regressor (SVR)** obteve um desempenho um pouco inferior (R^2 de 0.9675 e RMSE de 0.5992), mesmo após eu aplicar escalonamento nas features.

Considerações

- Todos os modelos com lag features superaram ou se igualaram ao desempenho do modelo baseline, o que reforça a importância dessas defasagens em séries temporais climáticas.
- A **Regressão Linear com lags**, apesar de ser um modelo simples, foi o que apresentou melhor resultado global (maior R^2 e menor RMSE).
- Os modelos baseados em árvores (Random Forest e Gradient Boosting) também se mostraram muito competentes, como esperado, já que pelo que eu entendi pela pesquisa, conseguem capturar relações não lineares e interações entre variáveis.

Conclusão

A inclusão de **lag features** foi uma estratégia muito boa para melhorar a previsão da temperatura média. Consegui melhorar bem o desempenho do modelo original de Regressão Linear, com resultados até melhores do que os modelos mais complexos como SVR ou Random Forest.