Previsão da temperatura horária para a cidade de São João del-Rei

Lucas Rômulo de Souza Resende[1]

[1]Universidade Federal de São João del-Rei (UFSJ)

Abstract. *The purpose of this work was to generate a machine learning model, using the recursive forecasting method on time series, seeking to perform an hourly temperature forecast in São João del-Rei's city. The model obtained manages to control the temperature forecasting with a MAE of 1.04, that is, it performs a forecast with an error of 1.04°C, and has an RMSE of 1.36, which indicates that there are no sudden variations in the predicted values.*

Resumo. *O intuito deste trabalho foi gerar um modelo de aprendizado de máquina, utilizando o método de predição recursiva em séries temporais, buscando realizar a previsão da temperatura horária no município de São João del-Rei. O modelo obtido consegue efetuar a predição da temperatura com um MAE de 1.04, ou seja, realiza previsões com erro de 1.04°C, e possui um RMSE de 1.36, o que indica que não existem variações bruscas nos valores preditos.*

1. Introdução

A previsão da temperatura é algo essencial na sociedade atual. Ela auxilia as pessoas em várias tomadas de decisões ao longo dos dias, como por exemplo: se é necessário sair com uma blusa a mais, levar guarda chuva e até mesmo saber se é um dia propício para um banho de sol.

Nesse projeto é tratada a previsão horária da temperatura na cidade de São João del-Rei a partir de dados obtidos do INMET, um instituto governamental. Esses dados são utilizados para o avaliação de modelos de aprendizado de máquina gerados a partir de duas abordagens: regressão e predição recursiva.

São gerados três modelos, dois recursores, utilizando algoritmos de *RandomForest* e *GradientBoosting*, e um modelo baseado em predição recursiva que utiliza o modelo gerado a partir do *RandomForest*, que se mostrou o melhor entre os dois recursores.

Entre os modelos obtidos, o que apresentou melhores resultados foi o que utiliza o método de predição recursiva, que conseguiu atingir um MAE de 1.04, RMSE de 1.36 e 4.8 de erro máximo.

Esses valores de erros indicam que o modelo consegue realizar a previsão da temperatura da cidade com um erro de ±1.04°C e com pouca variação, sendo necessário apenas o tratamento dos valores preditos para o fim da noite, quando ocorre o erro máximo, pois são observadas quedas bruscas de temperatura.

2. Metodologia

2.1. Dados

Os dados utilizados no projeto são dados climáticos obtidos através do site do INMET (Instituto Nacional de Meteorologia) e fazer referência a medições horárias de temperatura medidas em Graus Celsius (°C).

As medições são realizadas por uma estação automática que fica localizada no Campos Tancredo Neves (CTAN) próxima ao prédio da Ciência da Computação e da Moradia Estudantil, a uma altitude de 929.88 metros, nas coordenadas -21.106502, -44.250928. Sua posição pode ser observada no Google Maps.

Dentre os dados climáticos gerados pela estação, é possível observar informações como: data e hora da medição, precipitação total (chuva), pressão atmosférica, radiação global, temperatura do ar, temperatura do ponto de orvalho, umidade relativa do ar e direção e velocidade do vento.

Para o projeto foram selecionadas as variáveis: temperatura do ar, data e hora; que juntas representam a série temporal da temperatura.

Essa temperatura do ar é aferida utilizando um termômetro de bulbo seco, o tipo mais comum de termômetro, que não leva em consideração a umidade do ar no momento da medição da temperatura, e pode ser facilmente encontrado em vários locais, até mesmo em casas. Na Figura 1 pode ser visto um modelo simples de termômetro de bulbo seco.

Figure 1. Termômetro de parede

O intervalo de tempo dos dados utilizados nesse projeto são:

- Treino: de 01/jan/2013 a 31/dez/2022 (últimos 10 anos completos);
- Teste: de 01/jan/2023 a 10/jan/2023 (primeiros 10 dias de 2023).

A Figura 2 representa uma amostra retirada do dataset, onde é possível observar o target (temperatura) representado em valores reais utilizando ponto flutuante, a data da medição (data) representada no formato 'YY-mm-dd' e a hora da medição (hora) representada por um inteiro.

Novamente, na Figura 2, é possível perceber um valor '*NaN*' na coluna de temperatura, o que representa um valor nulo, sem medição. Esses valores correspondem a cerca de 4.38% das aferições de temperatura e foram tratados por meio de um método de imputação iterativa baseado nos trabalhos [Buck 1960] e [van Buuren and Groothuis-Oudshoorn 2011].

	temperatura	data	hora
39040	21.5	2017-06-15	16
65285	NaN	2020-06-13	5
84452	15.5	2022-08-20	20
27151	18.2	2016-02-06	7
41877	28.4	2017-10-11	21

Figure 2. Amostra do dataset

2.1.1. Análise dos Dados

A partir da análise do *box_plot* da temperatura, na Figura 3 é possível observar que a cidade de São João del-Rei possui uma temperatura amena que geralmente se encontra numa faixa de 16°C a 24°C. A cidade ainda possui uma mínima de aproximadamente 5°C e máxima próxima a 35°C.

As Figuras 4 e 5 apresentam, respectivamente, um histograma (distribuição da frequência) e a descrição estatística da variável temperatura.

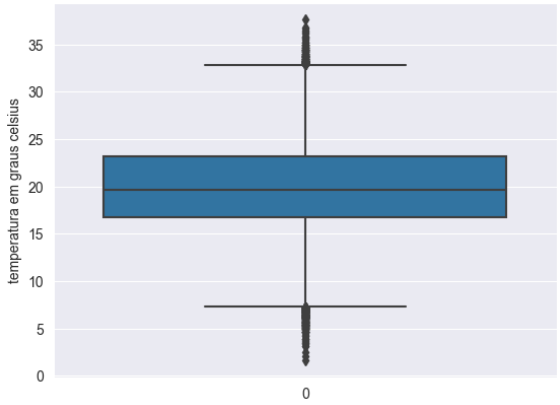

Figure 3. box_plot da temperatura

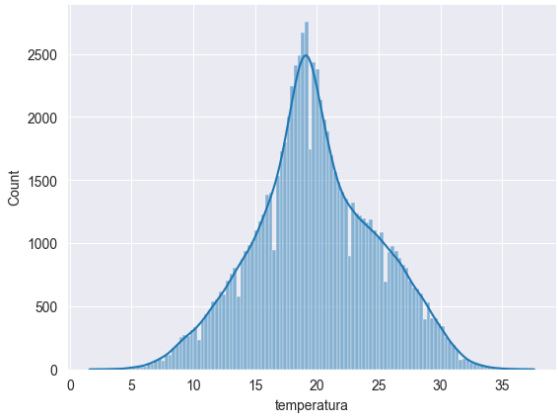

Figure 4. histograma da temperatura

	temperatura
count	83808.000000
mean	19.893916
std	4.972748
min	1.600000
25%	16.800000
50%	19.600000
75%	23.200000
max	37.600000

Figure 5. Descrição estatística da variável temperatura

Essas primeiras análises foram produzidas utilizando um dataset com os dados originais, ou seja, com cerca de 4.38% de dados faltantes. Esses dados faltantes foram corrigidos utilizando um método de imputação iterativa. A validação dessa correção se dá pela análise visual das Figuras 6, 7 e 8, onde é possível observar a manutenção das características dos dados.

Na Figura 6, que indica as médias horárias da temperatura de acordo com cada estação, é possível observar algumas leves alterações nas características dos dados, sendo a principal uma diminuição, de aproximadamente 1°C, da temperatura no verão no período entre 16h e 18h.

Na Figura 7, que indica as médias mensais da temperatura de acordo com cada estação, entende-se que as estações do ano em São João del-Rei são bem definidas: no verão a média da temperatura fica acima dos 22°C, no outono a média cai mês após mês até os 17°C, no primeiro mês do inverno ainda existe um decréscimo na temperatura e depois ela sobe até uma média de 20°C, e na primavera a média da temperatura se mantém entre os 20.5°C e 22°C. Também é possível observar a manutenção dessas características mesmo após a imputação dos dados faltantes.

Na Figura 8 é possível observar um gráfico de dispersão que possui cada uma das medidas horárias no intervalo entre 2013 e 2023.

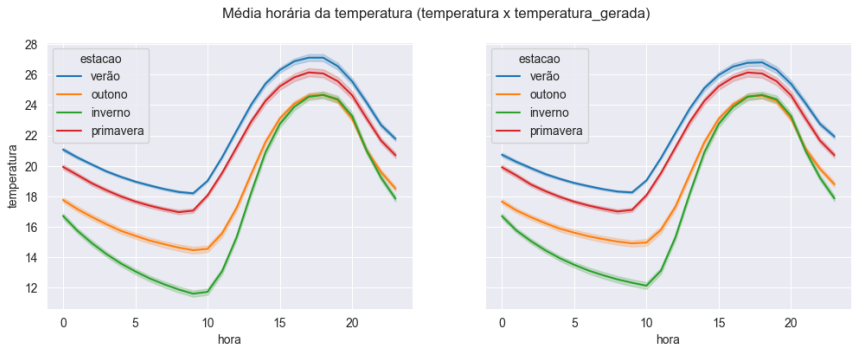

Figure 6. Comparação da média horária da temperatura

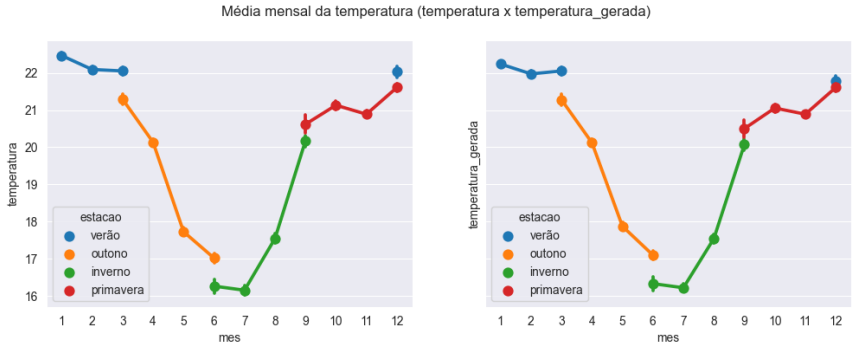

Figure 7. Comparação da média horária da temperatura

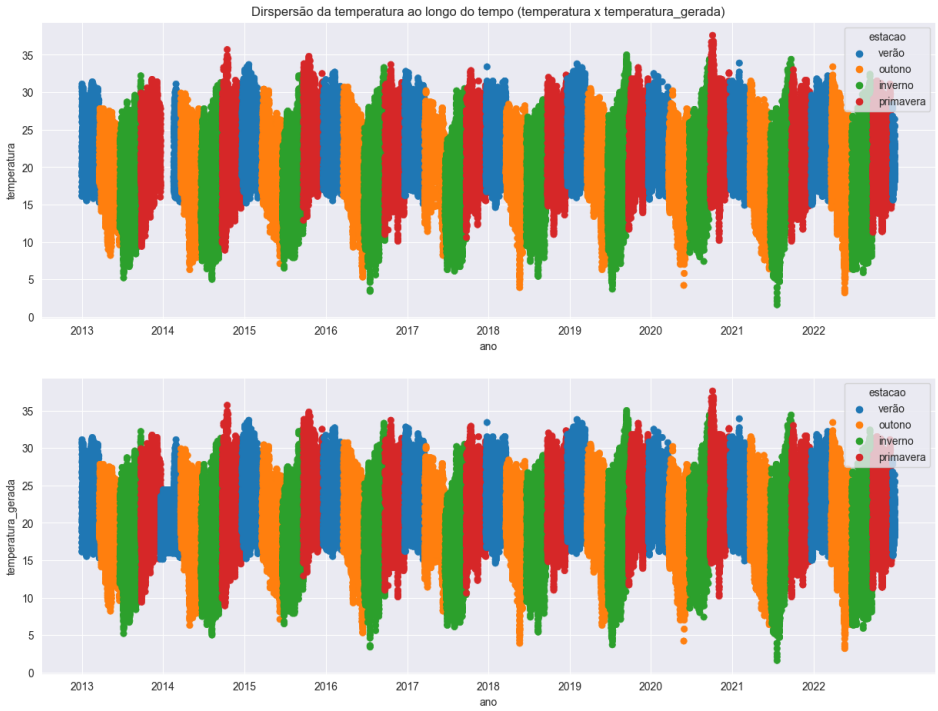

Figure 8. Comparação da dispersão da temperatura

É possível perceber uma faixa ausente de dados no fim do ano de 2014 e no início do ano de 2015, época do verão. Esses dados faltantes foram imputados, porém o *imputer* não conseguiu manter o aspecto da estação, agregando valores muito a baixo do esperado, o que, possivelmente, é o motivo das leves alterações nas características observadas nas Figuras 6 e 7.

Outros dados imputados não podem ser analisados visualmente, já que a quantidade de dados no gráfico é muito grande.

Um outro ponto importante a ser analisado são as tendências das temperaturas máximas e mínimas no decorrer dos anos, que podem ser vistas na Figura 9.

Esses aumentos na temperatura máxima e mínima é um assunto muito debatido entre pessoas das áreas de Ciências da Terra. No site da ONU é possível compreender como funcionam essas mudanças climáticas e de que forma a humanidade está contribuindo para a aceleração das mesmas.

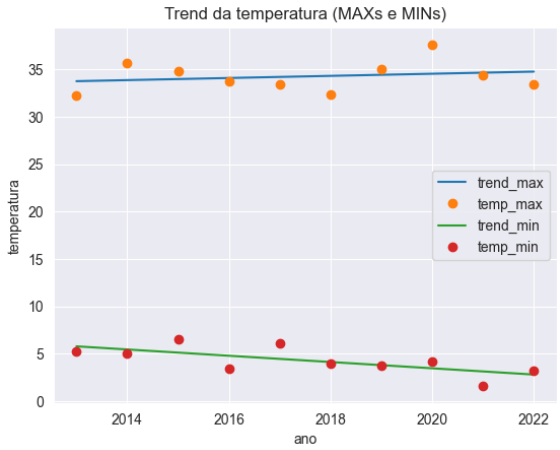

Figure 9. Tendência das máximas e mínimas

2.2. Machine Learning

Nessa etapa da metodologia foram treinados três modelos para a predição da temperatura da cidade de São João del-Rei. Dentre eles, dois modelos utilizam uma abordagem de regressão, o *GradientBoosting* e o *RandomForest*, e o outro modelo utiliza a abordagem de *RecursiveForecast*, ou *Predição Recursiva*.

2.2.1. Gradient Boosting

O *GradientBoosting*, descrito em [Friedman 2001] e [Friedman 1999], é um método que consiste na geração sequencial de estimadores (*DecisionTrees* ou *Stomps*), cada um treinado utilizando a perda calculada do estimador anterior, buscando ”consertar os erros” cometidos.

O modelo gerado utilizando o algoritmo de *GradientBoosting* que obteve menos erros foi configurado modificando o valor da taxa de aprendizado de .1 para .05, enquanto os outros parâmetros foram mantidos em seus valores padrão. Esse modelo obteve erros

de aproximadamente 2.2 de *MAE*, 2.89 de *RMSE* e um erro máximo de 7.71, sendo o pior entre os três modelos gerados.

A Figura 10 exibe as predições realizadas pelo modelo em comparação com os valores reais da temperatura.

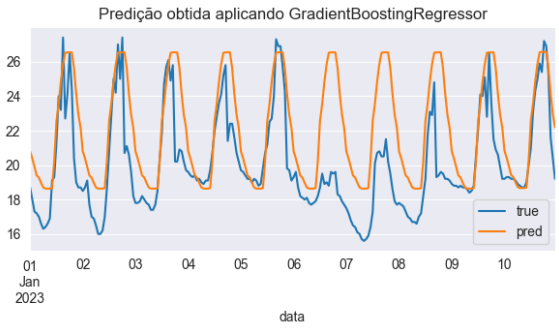

Figure 10. Aplicação do modelo do GradientBoosting

2.2.2. Random Forest

O *RandomForest*, descrito em [Breiman 2001] e [Geurts et al. 2006], é um método que consiste na geração de vários estimadores do tipo árvore de decisão (*DecisionTrees*) e, após a execução da regressão por esses estimadores, o resultado obtido é o resultado mais frequente entre os estimadores.

O modelo gerado utilizando o algoritmo *RandomForest* que obteve menos erros foi configurado modificando a quantidade de estimadores gerados de 100 para 150 e utilizando o critério '*friedman_mse*', enquanto os outros parâmetros foram mantidos em seus valores padrão. Esse modelo obteve erros de aproximadamente 2.2 de *MAE*, 2.89 de *RMSE* e um erro máximo de 7.71, sendo o pior entre os três modelos gerados.

da taxa de aprendizado de .1 para .05, enquanto os outros parâmetros foram mantidos em seus valores padrão. Esse modelo obteve erros de aproximadamente 1.9 de *MAE*, 2.43 de *RMSE* e um erro máximo de 7.3, sendo o melhor modelo dentre os regressores, porém obteve resultados piores do que o método de predição recursiva.

A Figura 11 exibe as predições realizadas pelo modelo em comparação com os valores reais da temperatura.

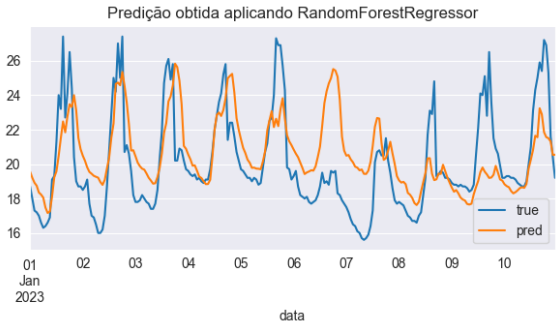

Figure 11. Aplicação do modelo do RandomForest

2.2.3. Recursive Forecasting

O método de *RecursiveForecasting* ou *Predição Recursiva*, consiste na utilização de um regressor para gerar predições sequencialmente, sempre utilizando as últimas predições geradas no cálculo das próximas.

O modelo gerado utiliza o modelo regressor do *RandomForest*, que exibiu os melhores resultados entre os regressores, para a realização do método de predição recursiva. Esse modelo obteve erros de aproximadamente 1.04 de *MAE*, 1.37 de *RMSE* e um erro máximo de 4.8, obtendo os melhores resultados dentre os três modelos gerados.

A Figura 12 exibe as predições realizadas pelo modelo em comparação com os valores reais da temperatura.

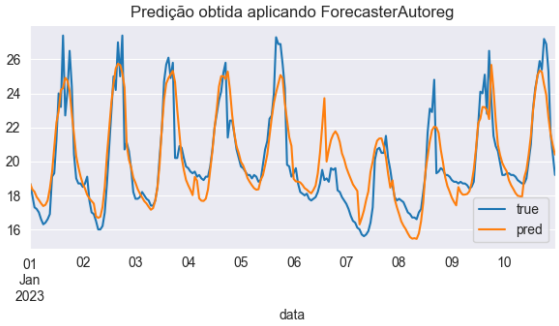

Figure 12. Aplicação do método *RecursiveForecasting*

2.2.4. Análises

Como exibido na Figura 13, entre os três modelos gerados (dois regressores e um preditor recursivo), o modelo que exibiu os melhores resultados foi aquele que utiliza do método de *Predição Recursiva*. O modelo que utiliza o *RandomForest* demonstrou ser pior que o preditor recursivo, porém ficou sendo o melhor dentre os regressores, enquanto o modelo que utiliza o algoritmo de *GradientBoosting* apresentou os piores resultados.

Os erros do melhor modelo, o preditor recursivo, foram de 1.04 de MAE, 1.37 de RMSE e 4.8 de erro máximo.

	model	mean_absolute_error	max_error	RMSE
0	GradientBoosting	2.198198	7.713930	2.887426
1	RandomForest	1.896953	7.305333	2.427012
2	Forecast	1.037904	4.816000	1.369398

Figure 13.

Esses erros indicam que as predições realizadas pelo modelo não possuem uma variação muito alta, visto que o RMSE é de menos de 1.37 e está bem próximo do MAE; e os valores de MAE e erro máximo representam, diretamente, os erros obtidos de temperatura. Sendo assim, a partir do MAE é possível compreender que o modelo consegue predizer valores para a temperatura com uma precisão de ±1.04°C.

O erro máximo do modelo, de 4.8°C, é um valor preocupante, pois essa variação na temperatura dessa cidade já indica um clima diferente. Entretanto, a partir de uma análise da Figura 12, é possível observar que esse erro ocorre próximo do fim da noite, quando a temperatura cai bruscamente, enquanto que o modelo realiza um queda mais suave. O reconhecimento desse fato facilita no tratamento desse erro.

3. Conclusão

Como observado na análise dos modelos gerados, quando se tratando de dados que representam uma série temporal, modelos de predição recursiva são mais eficazes do que modelos de regressão, apresentando RMSE de mais de 1 ponto de diferença.

A partir da utilização do método de predição recursiva é possível realizar a predição da temperatura da cidade de São João del-Rei com uma precisão de ±1.04°C, variação que não influencia tanto no clima e na sensação térmica, sendo então um bom modelo para a realização dessa tarefa.

References

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

[Buck 1960] Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(2):302–306.

[Friedman 1999] Friedman, J. (1999). Stochastic gradient boosting. department of statistics.

[Friedman 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[Geurts et al. 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

[van Buuren and Groothuis-Oudshoorn 2011] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.