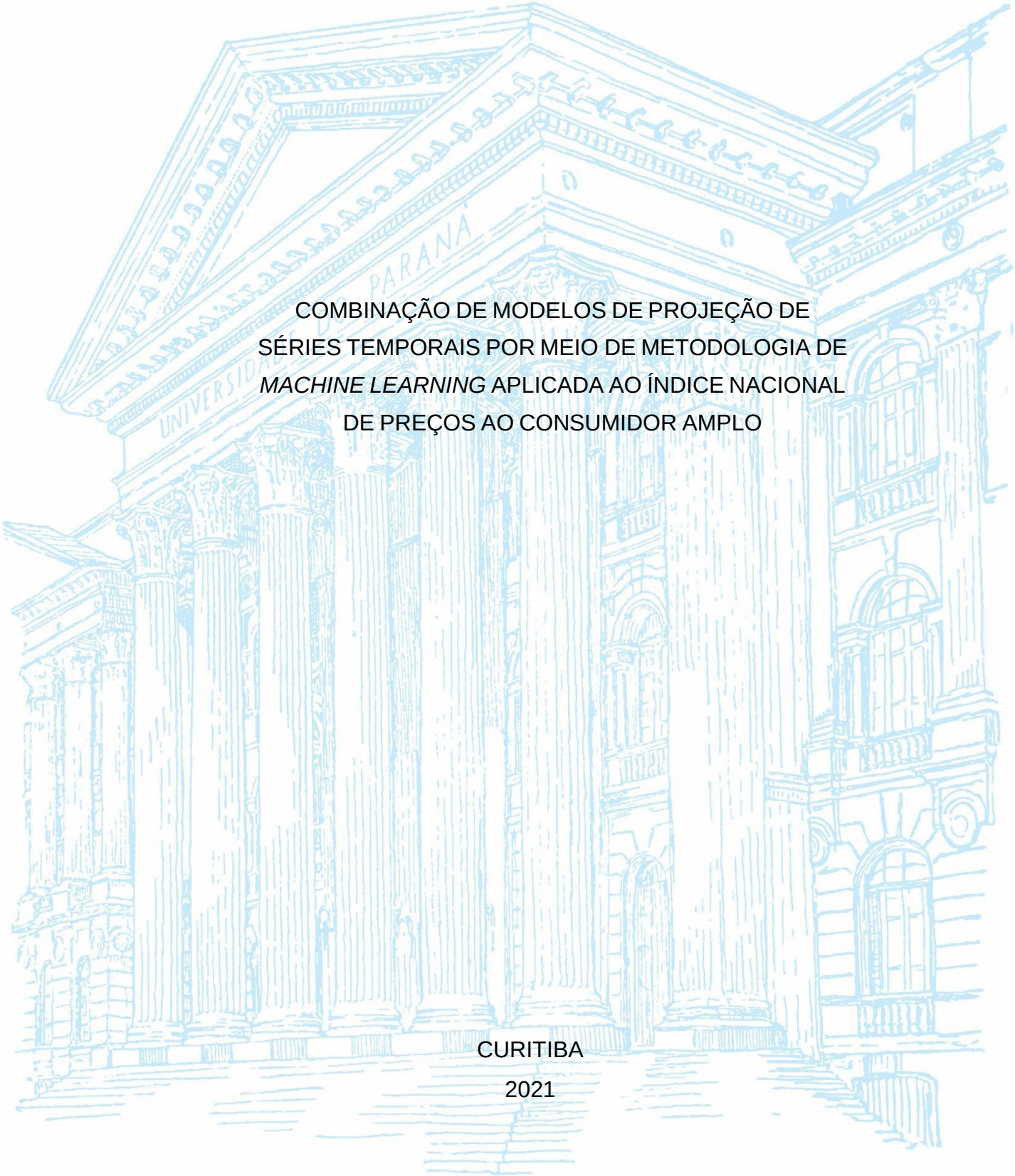
UNIVERSIDADE FEDERAL DO PARANÁ

GUSTAVO ELOI STABELINI NASCIMENTO



COMBINAÇÃO DE MODELOS DE PROJEÇÃO DE SÉRIES TEMPORAIS POR MEIO DE METODOLOGIA DE *MACHINE LEARNING* APLICADA AO ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR AMPLO

CURITIBA

2021

GUSTAVO ELOI STABELINI NASCIMENTO

COMBINAÇÃO DE MODELOS DE PROJEÇÃO DE SÉRIES TEMPORAIS POR MEIO DE METODOLOGIA DE *MACHINE LEARNING* APLICADA AO ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR AMPLO

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Produção da Universidade Federal do Paraná como requisito à obtenção do título de Bacharel em Engenharia de Produção.

Orientadora: Mariana Kleina

CURITIBA

2021

RESUMO

As séries temporais sempre trazem muitas informações em seu histórico, que podem ser utilizadas para projetar seu comportamento futuro. As projeções são extremamente importantes para as empresas, pois elas garantem que a demanda poderá ser suprida. Existem diversas formas de se realizar uma projeção e de medir sua acurácia, assim, este trabalho teve como objetivo avaliar a acurácia de diferentes métodos de projeção aplicadas ao Índice Nacional de Preços ao Consumidor Amplo (IPCA). Foram avaliados sete modelos de projeções de séries temporais, incluindo uma combinação dos outros seis modelos gerada por um algoritmo de *Machine Learning*, com a metodologia *Random Forest*.

Palavras-chaves: Séries Temporais. Projeções. Acurácia. Aprendizado de Máquina. Árvores de Decisão.

ABSTRACT

Time Series can bring a lot of important information in their history, that can be used to predict their future. Time Series Forecasts are extremely important to organizations, because they can ensure that the demand will be met. There are a lot of different ways to generate forecasts and measure their accuracy, therefore, the goal of this study is to evaluate the accuracy of different methods of forecasting applied to the IPCA. Seven forecasting methods were evaluated, including a combination of the other six by a Machine Learning algorithm, using the Random Forest technique.

Keywords: Time Series. Forecasting. Accuracy. Machine Learning. Random Forest.

SUMÁRIO

1 INTRODUÇÃO

No setor industrial, segundo Júnior (2007), um dos principais fatores a contribuir para a eficiência da cadeia produtiva das empresas que operam com ênfase na produção para estoque é a previsão da demanda, que é fundamental para o planejamento da produção, e também para o início do processo de suprimento. Assim, o desempenho da empresa está diretamente relacionado à acurácia de suas previsões. Pode-se perceber que grandes variações nesse processo implicam em não atendimento da demanda, atraso no atendimento e/ou excesso de estoques. Essas condições certamente aumentam custos e afetam a lucratividade do negócio.

Este trabalho apresenta um estudo sobre metodologias de projeções de séries temporais, que visa analisar a acurácia de diferentes modelos. Cada modelo será avaliado individualmente, e depois será feito um estudo utilizando o *Random Forest*, técnica de *Machine Learning*, que irá ponderar esses modelos, formando uma nova equação de projeção. O objetivo é comparar a acurácia dos modelos individuais com o modelo gerado a partir do Random Forest, para verificar a viabilidade da utilização do Machine Learning em projeções de séries temporais aplicando a série histórica do Índice Nacional de Preços ao Consumidor Amplo (IPCA).

2. REFERENCIAL TEÓRICO

2.1 SÉRIES TEMPORAIS

Uma série temporal pode ser definida como uma coleção de observações feita sequencialmente ao longo do tempo, geralmente em intervalos uniformes. São exemplos de séries temporais: Vendas diárias de jornal em uma banca, população de um determinado país ao longo dos anos e temperatura (ºC) de uma cidade ao longo do dia.

A maioria das séries temporais são estocásticas, ou seja, podem ser definidas por um componente determinístico (que pode ser modulado) e um componente estocástico (aleatório).

Ehlers (2007) diz que um processo estocástico pode ser definido matematicamente como uma coleção de variáveis aleatórias ordenadas no tempo e definidas em um conjunto de pontos.

As séries temporais apresentam algumas características, que serão definidas a seguir.

2.1.1 Tendência

Padrão de crescimento ou decrescimento da variável ao longo do tempo.

2.1.2 Sazonalidade

Variações relacionadas a algum período de tempo, como semana, mês, ano, etc.

2.1.3 Ciclos

Variações que, apesar de periódicas, não são associadas a nenhuma medida temporal. São exemplos: Ciclos econômicos e ciclos de epidemias.

2.1.4 Estacionariedade

Uma série estacionária é aquela que se desenvolve com média, variância e autocorrelação constantes. Isso implica em um equilíbrio, ou estabilidade estatística nos dados. (MONTGOMERY & JENNINGS, 2008).

2.1.5 Autocorrelação

A autocorrelação tem como objetivo verificar se as observações são dependentes e estão relacionadas entre si na série, o que contribui para identificar padrões como a sazonalidade. É obtida por meio do coeficiente de autocorrelação (ACF – *Autocorrelation Function*), onde a série é defasada em k períodos (ou *lags*) e comparada com ela mesma.

Montgomery e Jennings (2008) definem o ACF com uma defasagem k para uma série estacionária como:

$$\rho_k = \frac{E[(y_t - \mu)(y_{t+k} - \mu)]}{\sqrt{E[(y_t - \mu)^2]E[(y_{t+k} - \mu)^2]}} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\gamma_t}{\gamma_0}$$

Onde:

y_t = observação no tempo t,

μ = média das observações,

γ_t = função de autocovariância no tempo t,

E = esperança matemática,

Cov = covariância,

Var = variância.

2.2 PROJEÇÃO

Um dos principais objetivos ao analisar séries temporais é realizar projeções, que consiste em prever o comportamento esperado da série, com base em seu histórico, por meio de procedimentos matemáticos. Alguns dos métodos mais conhecidos serão abordados neste trabalho.

Para Wheelwright (1985), os métodos de previsão supõem que as observações passadas (histórico de dados) contêm todas as informações sobre o comportamento

da série. Eles extrapolam as características observadas e geram previsões confiáveis se o futuro apresentar comportamento similar ao passado.

2.2.1 MÉDIA MÓVEL SIMPLES (MMS)

Este método, também conhecido como *Moving Average*, consiste em aplicar a média aritmética dos *n* últimos períodos da demanda observada. (PEINADO & GRAEML, 2007).

$$Y_t = \frac{\sum_{i=0}^{n} Y_{t-i}}{n} \, , n > 0,$$

Onde:

Y_t = Previsão no tempo t,

n = número de períodos utilizados para apurar a média móvel.

Quanto maior for o valor de *n,* maior será a influência das observações mais antigas. Na prática seu valor geralmente é 3, ou seja, a projeção para o próximo período, nesse caso, seria a média das últimas 3 observações.

2.2.2 MÉTODO NAÏVE

Esse é um método muito simples, que consiste apenas em aplicar o último valor observado para todos os períodos de projeção.

$$Y_t = Y_{t-1},$$

Onde:

Y_t = Previsão no tempo t.

Esse modelo é muito eficiente para séries temporais utilizadas nas áreas de finanças e economia. (HYNDMAN & ATHANASOPOULOS, 2018).

2.2.3 MÉTODO DRIFT

Esse método pode ser considerado uma alternativa para o método Naïve, com uma variação que permite que a projeção cresça ou decresça com o tempo. Essa

variação corresponde à média vista no histórico dos dados. (HYNDMAN & ATHANASOPOULOS, 2018).

$$Y_{T+h} = Y_T + \frac{h}{T-1}\sum_{t=2}^{T}(Y_t - Y_{t-1}) = Y_T + h\left(\frac{Y_T - Y_1}{T-1}\right),$$

Onde:

Y_{T+h} = Previsão no tempo $T+h$.

Isso equivale a desenhar uma reta do primeiro ao último ponto e extrapolar esta reta para o futuro.

2.2.4 MÉTODO DE HOLT

Método proposto por Holt em 1957 que vai além do alisamento exponencial simples, proposto anteriormente, considerando agora a tendência dos dados. (HYNDMAN & ATHANASOPOULOS, 2018).

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}), 0 \leq \alpha \leq 1,$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, 0 \leq \beta \leq 1,$$

Onde:

l_t = equação de nível no tempo t,

b_t = equação de tendência no tempo t,

α = coeficiente de nível,

β = coeficiente de tendência.

A equação de projeção se dá por:

$$Y_{t+h} = l_t + hb_t.$$

Existem diferentes formas de se estimar os valores para os coeficientes α e β. Uma forma muito comum é utilizar o Erro Médio Absoluto, que será abordado futuramente neste trabalho.

2.2.5 MÉTODO DE HOLT WINTERS

Esse método pode ser considerado uma extensão do método de Holt, agora considerando também a sazonalidade. Existem duas variações desse modelo, que se diferem pela natureza do componente sazonal. (HYNDMAN & ATHANASOPOULOS,

2018).

Se sua variação for multiplicativa, ou seja, com amplitude que varia com o tempo, tem-se:

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \, , 0 \leq \alpha \leq 1,$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} \, , 0 \leq \beta \leq 1,$$
$$s_t = \gamma \frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma) s_{t-m} \, , 0 \leq \gamma \leq 1,$$

Onde:

l_t = equação de nível no tempo t,

b_t = equação de tendência no tempo t,

s_t = equação de sazonalidade no tempo t,

α = coeficiente de nível,

β = coeficiente de tendência,

γ = coeficiente de sazonalidade,

m = frequência da sazonalidade.

A equação de projeção se dá por:

$$Y_{t+h} = (l_t + hb_t) s_{t+h-m}.$$

Já no caso de sazonalidade aditiva, quando as variações sazonais são aproximadamente constantes, tem-se:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \, , 0 \leq \alpha \leq 1,$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} \, , 0 \leq \beta \leq 1,$$
$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma) s_{t-m} \, , 0 \leq \gamma \leq 1,$$

A equação de projeção se dá por:

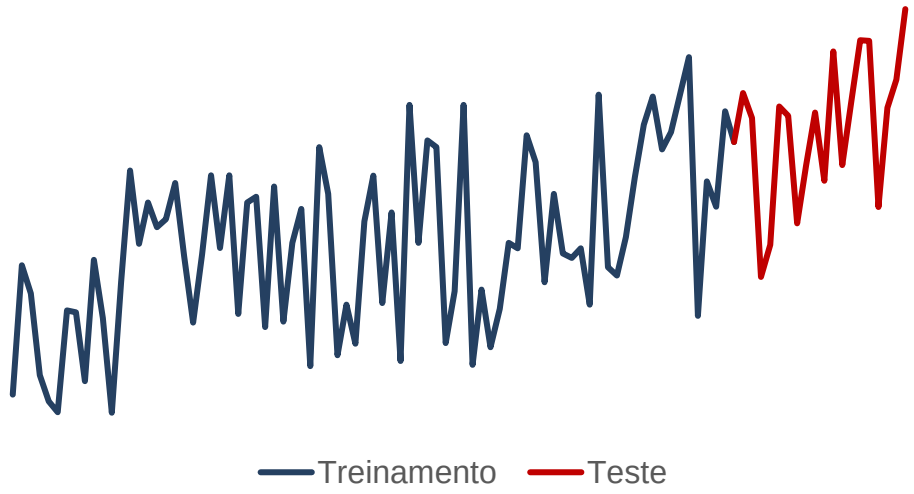
$$Y_{t+h} = l_t + hb_t + s_{t+h-m}.$$

2.3 AVALIANDO MÉTODOS E ESTIMANDO PARÂMETROS

Nesse tópico, serão abordadas algumas das formas mais comuns utilizadas para escolher o modelo final de projeção e também como otimizar os valores dos parâmetros dos métodos quando necessário.

Primeiramente, os dados são divididos em dois grupos, conforme Figura 1: os dados que treinam e os que testam os métodos de projeção.

FIGURA 1: SEPARAÇÃO DOS DADOS



FONTE: Autor (2020)

É muito comum utilizar 20% das observações como teste, mas esse valor depende do tamanho da amostra e também de quantas observações serão projetadas.

Existem diferentes formas de se medir o erro de projeção, que serão abordadas a seguir, o importante é entender que tanto para otimizar os valores dos parâmetros dos modelos quanto para tomar a decisão de qual método utilizar, o objetivo será minimizar o erro.

Segundo Hyndman e Athanasopoulos (2018), também é importante ter em mente que a acurácia das projeções só poderá ser determinada considerando como o modelo conseguiu prever os novos valores, sem considerar os dados utilizados para ajustar os modelos. Um modelo bem ajustado aos dados de teste não irá necessariamente gerar uma boa projeção.

2.3.1 ERROS DE PROJEÇÃO

O erro de projeção é definido como a diferença entre o valor observado e a projeção feita para o mesmo período, e pode ser escrito da seguinte forma:

$$e_t = y_t - \hat{y}_t,$$

Onde:

e_t = erro da projeção no tempo t,

y_t = valor observado no tempo t,

$\hat{y}_t$ = projeção para o tempo t.

2.3.1.1 Erro Médio Absoluto (*Mean Absolute Error – MAE*)

O MAE é equivalente à média do módulo dos erros para todos os períodos projetados.

$$MAE = média(|e_t|).$$

2.3.1.2 Erro Quadrático Médio (*Mean Squared Error – MSE*)

O MSE é equivalente à média dos quadrados dos erros para todos os períodos projetados.

$$MSE = média(e_t^2).$$

2.3.1.3 Raiz do Erro Quadrático Médio (*Root Mean Squared Error - RMSE*)

Para que os valores estejam na mesma unidade, o RMSE se equivale à raiz quadrada do MSE.

$$RMSE = \sqrt{MSE}.$$

2.4 *MACHINE LEARNING*

Aprendizado de máquina (*machine learning*), segundo Corrêa (2014), é a ciência de fazer com que os computadores aprendam a agir como os humanos, e melhorem seu aprendizado ao longo do tempo de maneira autônoma, alimentando-se de dados e informações na forma de observações e interações do mundo real. Esse conhecimento adquirido permite que os computadores generalizem seu conhecimento corretamente para novas configurações e situações.

2.5 *RANDOM FOREST*

2.5.1 Árvores de Decisão

Segundo Garcia (2000), árvores de decisão são uma forma mais simples e eficaz de representar o conhecimento. Elas baseiam-se na abordagem “dividir para conquistar”, ou seja, na sucessiva divisão do conjunto de exemplos utilizado para o treino, em vários subconjuntos, até cada um destes subconjuntos pertencer a uma mesma classe, ou até uma das classes ser majoritária, não havendo necessidade de novas divisões.

Os resultados dos vários subconjuntos obtidos com a construção de uma árvore de decisão são dados organizados de maneira compacta, utilizados para classificar novos exemplos.

2.5.2 Algoritmo Random Forest

O método conhecido como Random Forest é um tipo de *ensemble learning* (aprendizado em conjunto), que gera muitos classificadores e combina seu resultado.

O algoritmo foi desenvolvido por Leo Breiman e Adele Cutler (BREIMAN, 2001).

Segundo Cutler, Cutler e Stevens (2012), tem-se que:

Como o nome sugere, *Random Forest* é um método de aprendizado em conjunto baseado em árvores de decisão, onde cada árvore depende de uma coleção de variáveis aleatórias. Mais formalmente, para um vetor aleatório com dimensão *p (* $X = (X_1, \dots, X_p)^T$), representando os valores reais de entrada, e uma variável aleatória Y, representando o valor real de resposta, assume-se uma distribuição conjunta desconhecida $P_{XY}(X, Y)$. O objetivo é encontrar uma função $f(X)$ para prever Y. A função preditiva é determinada por uma função de perda $L(Y, f(X))$ e é definida para minimizar o valor esperado do erro:

$$E_{XY}(L\big(Y, f(X)\big)),$$

onde os subscritos denotam o valor esperado em relação à distribuição conjunta de X e Y.

Intuitivamente, $L(Y, f(X))$ é uma medida de quão próximo $f(X)$ está de Y. Assim, valores de $f(X)$ muito distantes de Y são penalizados. Os tipos de erros mais comuns utilizados são o erro quadrático $L\big(Y, f(X)\big) = (Y - f(X))^2$ para problemas de regressão, e o erro *“zero-one”* para classificação:

$$L\big(Y, f(X)\big) = I\big(Y \neq f(X)\big) = \begin{cases} 0, se\ Y = f(X) \\ 1, caso\ contrário \end{cases}.$$

Acontece que, minimizando $E_{XY}(L\big(Y, f(X)\big))$ para um erro quadrático, obtém-se a função:

$$f(X) = E(X = x),$$

conhecida como função de regressão. No caso da classificação, se os possíveis valores de Y são denotados por ч, minimizando $E_{XY}(L\big(Y, f(X)\big))$ para erros do tipo *"zero-one"* tem-se:

$$f(X) = argP(X = x),$$

conhecida como regra de Bayes.

Algoritmos *ensembles* produzem f em termos de uma coleção das chamadas "bases de aprendizado" $h_1(x), \dots, h_J(x)$ e essas bases são combinadas para gerar o "preditor em conjunto" $f(x)$. Na regressão, as bases de aprendizado têm média

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x),$$

quando em classificação, $f(x)$ é a classe mais frequentemente prevista.

$$f(x) = arg\sum_{j=1}^{J} I(y = h_j\,(x)).$$

No *Random Forest*, a j-ésima base de aprendizado é uma árvore denotada como $h_j(X, \Theta_j)$, onde Θ_j é uma coleção de variáveis aleatórias e os Θ_j's são independentes para: $j = 1, \dots, J$. (CUTLER; CUTLER; STEVENS, 2012).

2.6 IPCA

O Índice Nacional de Preços ao Consumidor Amplo (IPCA) é um índice gerado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que mede a variação de preços de mercado para o consumidor, e funciona como um termômetro para a inflação no país.

Segundo o próprio IBGE (2020),

> “O IPC é calculado como uma média ponderada das variações de preços dos bens e serviços que integram uma cesta fixa coberta pelo índice. Nessa média, os pesos devem refletir a importância relativa dos produtos na cesta, medida pela participação de cada um deles na despesa de consumo total das famílias. Os pesos associados a cada produto determinam o grau de influência que seu movimento de preços terá sobre o índice geral, e devem retratar os hábitos e o perfil de consumo médio da população coberta pelo índice” (IBGE, 2020).

Ainda segundo o IBGE, atualmente, a população-objetivo do IPCA abrange as famílias com rendimentos de 1 a 40 salários mínimos, qualquer que seja a fonte, residentes nas áreas urbanas das regiões de abrangência do SNIPC, as quais são: regiões metropolitanas de Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Vitória, Rio de Janeiro, São Paulo, Curitiba, Porto Alegre, além do Distrito Federal e dos municípios de Goiânia, Campo Grande, Rio Branco, São Luís e Aracaju.

Para a montagem das estruturas de ponderadores é utilizada uma organização de códigos em grupamentos estabelecidos de forma lógica, fazendo com que algumas categorias específicas permaneçam juntas, como por exemplo, categorias de consumo de mesma natureza, hierarquicamente estruturadas em grupos, subgrupos, itens e subitens. Os pesos utilizados no cálculo dos índices de preços são obtidos pelo nível mais desagregado, representado pelos grupos, subgrupos, itens e subitens anteriormente mencionados. Esses ponderadores apresentam o grau de importância ou representatividade dos subitens pertencentes à cesta de consumo das famílias, que são elaboradas a partir dos hábitos de consumo da população-alvo da pesquisa.

Com relação à base de dados, a série histórica está posicionada em dezembro de 1993, que possui valor igual a 100 (base = 100). Os valores são atualizados todos os meses e dizem respeito à variação mensal nos preços analisados.

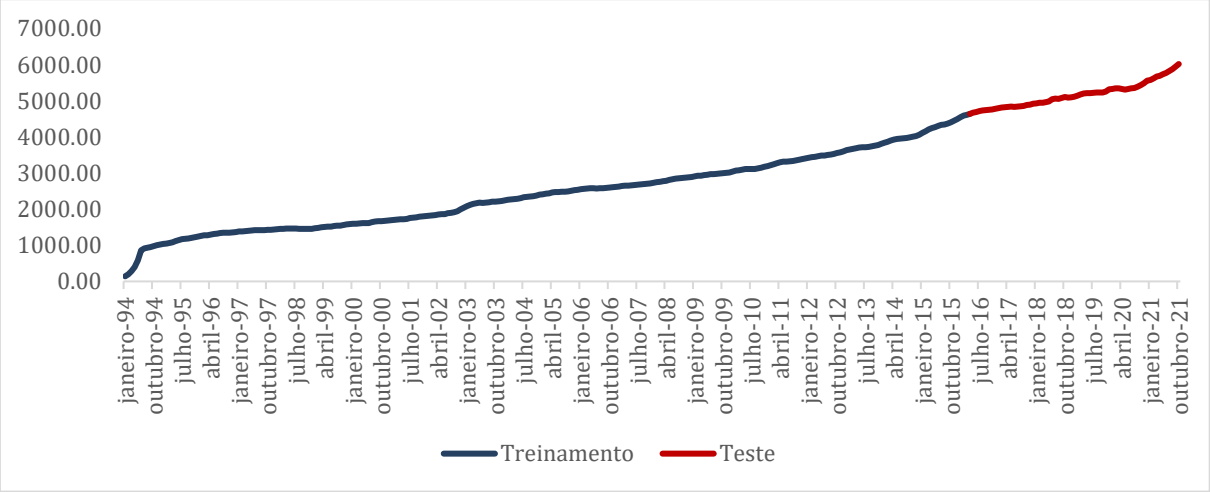
3 METODOLOGIA

3.1 PROJEÇÕES

Para realizar as projeções, o software utilizado foi o R, por ser um software muito utilizado no mundo todo para fins como esse. Nele se encontram muitas bibliotecas que facilitam a geração das projeções. Neste trabalho, foi utilizada a biblioteca “fpp2”, que se encontra no livro “Forecasting: principles and practice”, que pode ser encontrado nas Referências deste trabalho.

Na série utilizada, tem-se um total de 334 observações (Janeiro 1994 até Outubro 2021). Dividindo essas observações com uma proporção de 80%-20%, tem-se 267+67 observações, como se vê na Figura 2, onde o eixo X representa as datas a partir de Janeiro de 1994 e o eixo Y representa o valor do IPCA. As primeiras 267 observações (Treinamento) foram utilizadas para as projeções iniciais.

FIGURA 2: DADOS DA SÉRIE TEMPORAL DO IPCA



FONTE: Autor (2021)

Na Figura 3, tem-se um exemplo da utilização da biblioteca “fpp2” na projeção inicial com o método de Holt:

FIGURA 3: EXEMPLO DE PROJEÇÃO UTILIZANDO O R

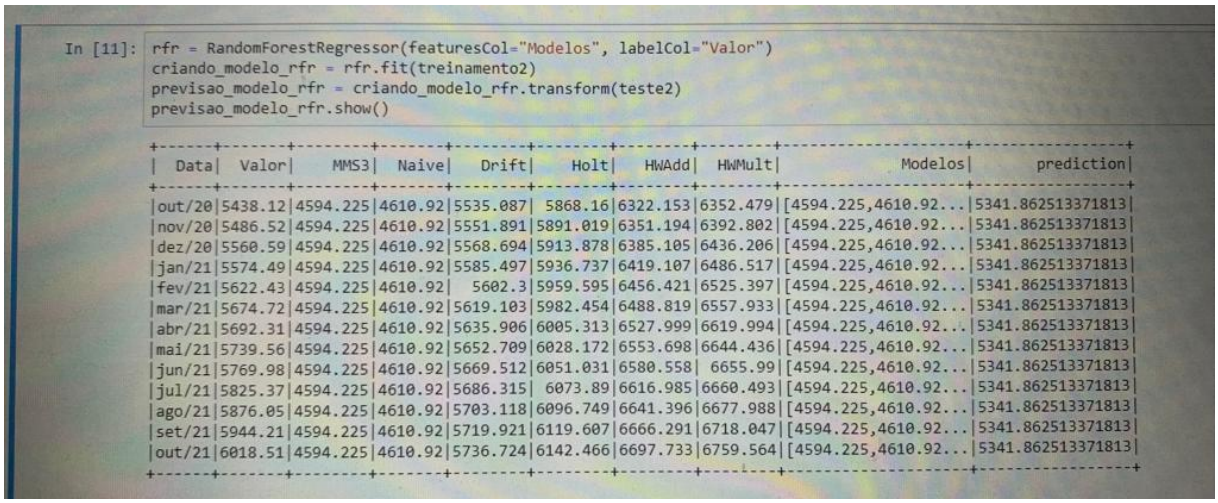
```
> holt <- c(valores[1:267])
> library(fpp2)
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
-- Attaching packages --------------------------------------
v ggplot2   3.3.5     v fma       2.4
v forecast  8.15      v expsmooth 2.3

> holt(holt,67)
    Point Forecast      Lo 80     Hi 80       Lo 95     Hi 95
268       4633.781 4611.46032  4656.102  4599.64437  4667.918
269       4656.640 4609.40553  4703.875  4584.40109  4728.879
270       4679.499 4602.39879  4756.599  4561.58446  4797.413
271       4702.358 4591.07481  4813.641  4532.16516  4872.550
272       4725.217 4575.90085  4874.532  4496.85785  4953.575
273       4748.076 4557.22468  4938.926  4456.19435  5039.957
274       4770.934 4535.31584  5006.553  4410.58692  5131.282
275       4793.793 4510.39059  5077.196  4360.36629  5227.220
276       4816.652 4482.62720  5150.677  4305.80510  5327.499
277       4839.511 4452.17584  5226.846  4247.13300  5431.889
278       4862.370 4419.16518  5305.574  4184.54681  5540.193
279       4885.229 4383.70710  5386.750  4118.21761  5652.240
280       4908.088 4345.90002  5470.275  4048.29591  5767.879
281       4930.946 4305.83138  5556.061  3974.91546  5886.977
282       4953.805 4263.57955  5644.031  3898.19612  6009.414
283       4976.664 4219.21530  5734.113  3818.24609  6135.082
284       4999.523 4172.80292  5826.243  3735.16373  6263.882
285       5022.382 4124.40120  5920.363  3649.03894  6395.725
286       5045.241 4074.06413  6016.417  3559.95428  6530.527
```

FONTE: Autor (2021)

3.2 COMBINANDO OS MODELOS COM *RANDOM FOREST*

Para combinar os modelos, o software utilizado foi o Python, por ser um software utilizado com muita frequência quando se fala em *Machine Learning*, com muitas bibliotecas e cursos disponíveis na internet. Na Figura 4, é possível observar como a função *"RandomForestRegressor"* foi utilizada:

FIGURA 4: CÓDIGO DO RANDOM FOREST NO PYTHON



```
In [11]: rfr = RandomForestRegressor(featuresCol="Modelos", labelCol="Valor")
         criando_modelo_rfr = rfr.fit(treinamento2)
         previsao_modelo_rfr = criando_modelo_rfr.transform(teste2)
         previsao_modelo_rfr.show()
```

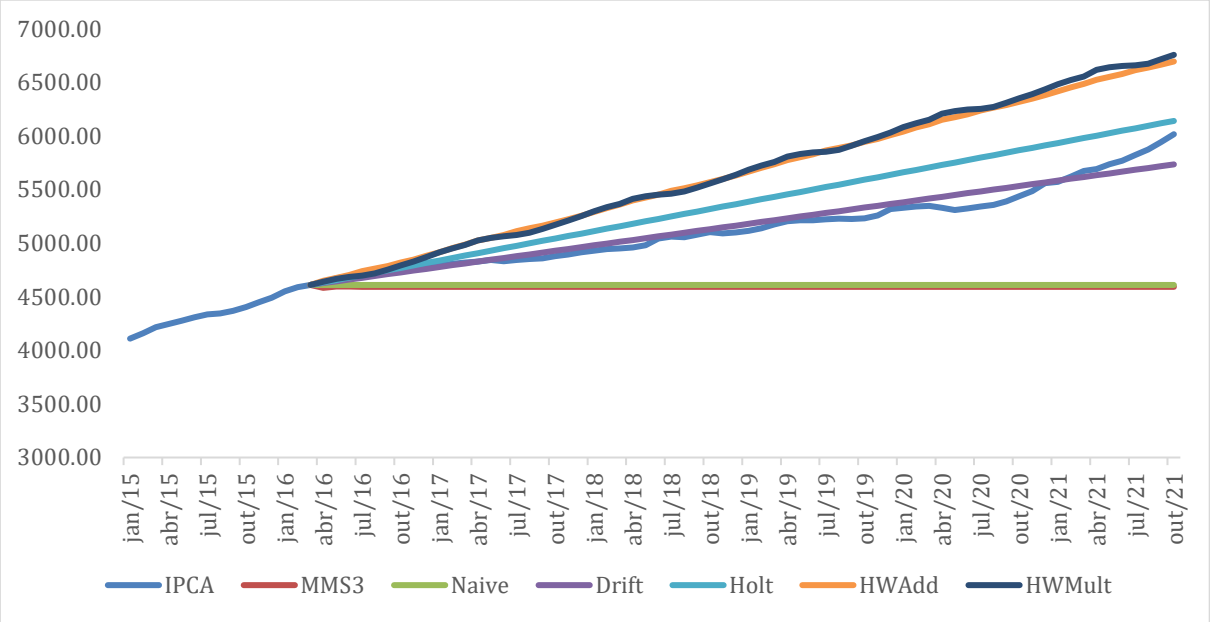
Data	Valor	MMS3	Naive	Drift	Holt	HWAdd	HWMult	Modelos	prediction
out/20	5438.12	4594.225	4610.92	5535.087	5868.16	6322.153	6352.479	[4594.225,4610.92...	5341.862513371813
nov/20	5486.52	4594.225	4610.92	5551.891	5891.019	6351.194	6392.802	[4594.225,4610.92...	5341.862513371813
dez/20	5560.59	4594.225	4610.92	5568.694	5913.878	6385.105	6436.206	[4594.225,4610.92...	5341.862513371813
jan/21	5574.49	4594.225	4610.92	5585.497	5936.737	6419.107	6486.517	[4594.225,4610.92...	5341.862513371813
fev/21	5622.43	4594.225	4610.92	5602.3	5959.595	6456.421	6525.397	[4594.225,4610.92...	5341.862513371813
mar/21	5674.72	4594.225	4610.92	5619.103	5982.454	6488.819	6557.933	[4594.225,4610.92...	5341.862513371813
abr/21	5692.31	4594.225	4610.92	5635.906	6005.313	6527.999	6619.994	[4594.225,4610.92...	5341.862513371813
mai/21	5739.56	4594.225	4610.92	5652.709	6028.172	6553.698	6644.436	[4594.225,4610.92...	5341.862513371813
jun/21	5769.98	4594.225	4610.92	5669.512	6051.031	6580.558	6655.99	[4594.225,4610.92...	5341.862513371813
jul/21	5825.37	4594.225	4610.92	5686.315	6073.89	6616.985	6660.493	[4594.225,4610.92...	5341.862513371813
ago/21	5876.05	4594.225	4610.92	5703.118	6096.749	6641.396	6677.988	[4594.225,4610.92...	5341.862513371813
set/21	5944.21	4594.225	4610.92	5719.921	6119.607	6666.291	6718.047	[4594.225,4610.92...	5341.862513371813
out/21	6018.51	4594.225	4610.92	5736.724	6142.466	6697.733	6759.564	[4594.225,4610.92...	5341.862513371813

FONTE: Autor (2021)

4 RESULTADOS

Após dividir os dados em 2 grupos, os dados de treinamento foram utilizados para gerar as projeções iniciais no R. Os resultados podem ser observados na Figura 5, onde o eixo X representa as datas a partir de Janeiro de 2015 e o eixo Y representa o valor do IPCA:

FIGURA 5: RESULTADO DAS PROJEÇÕES INICIAIS

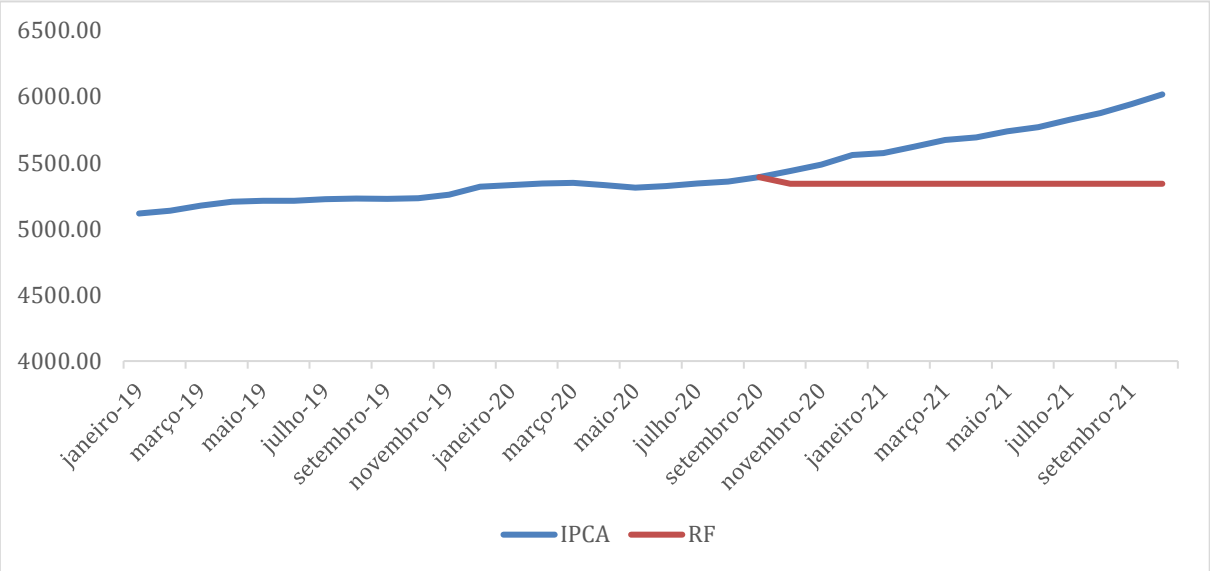


FONTE: Autor (2021)

Posteriormente, os dados das projeções iniciais (67 observações) foram separados em Treinamento (54 observações) e Teste (13 observações). Com isso, as primeiras 54 observações das projeções são combinadas, procurando se ajustar aos valores observados na série do IPCA, como se atribuíssemos pesos para os modelos, sendo que pesos maiores são dados aos modelos que mais se aproximam ao IPCA. Nas 13 restantes, os modelos são combinados da mesma forma que eles foram combinados no treinamento, e assim temos a projeção final com o *Random Forest*.

Na Figura 6, tem-se a projeção final realizada com o *Random Forest*, onde o eixo X representa as datas a partir de Janeiro de 2019 e o eixo Y representa o valor do IPCA:

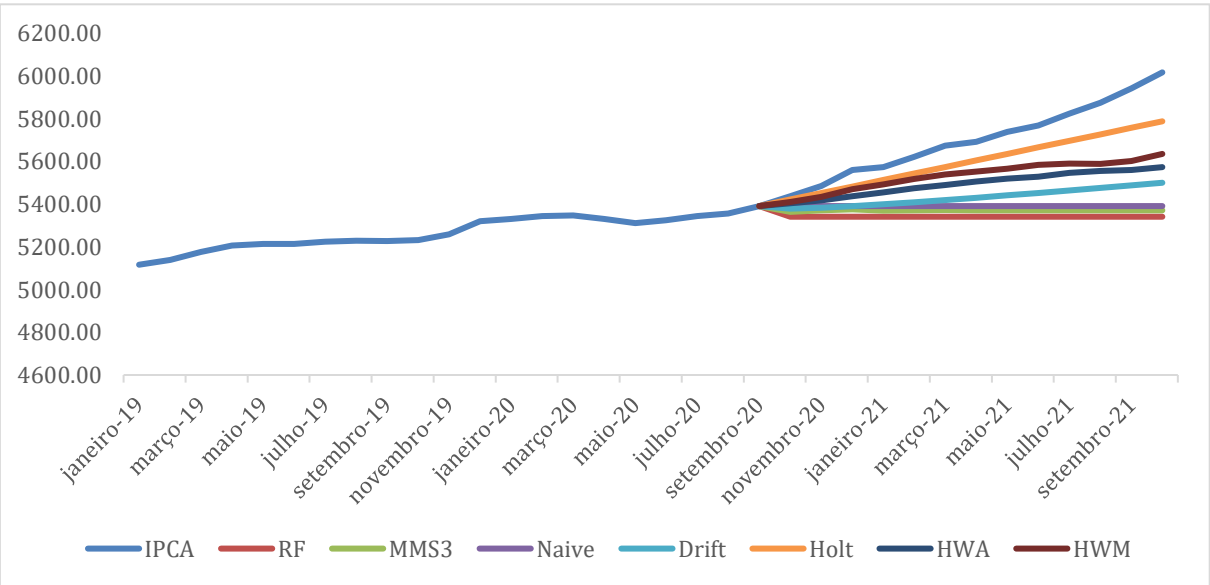
FIGURA 6: RESULTADO DA PROJEÇÃO COM O RANDOM FOREST



FONTE: Autor (2021)

Para podermos comparar de maneira justa as projeções realizadas, as projeções foram realizadas novamente a partir do mesmo ponto em que a projeção do Random Forest foi realizada. Assim, as projeções finais representam os últimos 13 meses observados. Na Figura 7, tem-se as projeções finais, incluindo a projeção realizada com o Random Forest, onde o eixo X representa as datas a partir de Janeiro de 2019 e o eixo Y representa o valor do IPCA:

FIGURA 7: RESULTADO DAS PROJEÇÕES FINAIS



FONTE: Autor (2021)

Para comparar os diferentes modelos, foi utilizada a Raiz do Erro Quadrático Médio (*Root Mean Squared Error - RMSE*), classificando-os do menor erro ao maior. Na Figura 8, pode-se observar os métodos e seus respectivos erros:

FIGURA 8: ERROS DOS MÉTODOS DE PROJEÇÃO

Método	Erro (RMSE)
Holt	118.06
HWM	201.82
HWA	240.79
Drift	304.30
Naive	359.63
MMS3	376.84
RF	404.37

FONTE: Autor (2021)

Para a combinação dos modelos de projeção com *Machine Learning*, foram realizados vários testes diferentes, visto que o modelo gerado não apresentou bons resultados. Como a projeção final foi uma reta (valor constante), o primeiro teste foi retirar da combinação os modelos de Média Móvel e o método Naive, que pelas suas fórmulas, sempre geram valores constantes. Sem sucesso, diferentes combinações foram feitas, utilizando os modelos descritos anteriormente e também modelos diferentes, mas sem bons resultados. Também foi feito um teste com outra linguagem de programação, o SAS. No SAS, foram testados diferentes modelos de projeção e também uma alternativa para o *Random Forest*, chamada *Gradient Boosting*, que é uma técnica mais complexa de *Machine Learning*, onde diversos parâmetros da função do *Machine Learning* foram alterados e até testes com outras séries temporais. Mesmo assim os resultados continuaram não apresentando ganho comparando-os com os modelos separadamente. O que foi observado é que a combinação dos modelos não gerou bons resultados para séries com tendência de crescimento. No fim, a decisão foi de continuar com a série do IPCA e a combinação utilizando todos os modelos descritos anteriormente. Assim, fica a sugestão para um trabalho futuro que busque entender o motivo disso ter acontecido.

5 CONCLUSÕES

Como já foi dito anteriormente, observou-se que a combinação dos modelos foi menos aderente aos dados da série do IPCA do que os próprios modelos utilizados, e futuros testes podem ser realizados em séries sem tendência de crescimento para obter melhores resultados.

Para este trabalho, foi mantida a proposta original de combinar os seis modelos descritos com a metodologia do Random Forest.

Primeiramente foram realizadas projeções iniciais utilizando o histórico da série do IPCA (primeiras 267 observações) com modelos de projeções de séries temporais. As 67 observações restantes foram divididas em 54 observações que foram treinadas por um algoritmo de *Machine Learning*, que combinou os modelos e 13 observações onde o algoritmo gerou uma nova projeção. Por fim, as projeções com os modelos de séries temporais foram realizadas novamente para as últimas 13 observações, para que pudessem ser comparadas.

O método de projeção que melhor se ajustou aos dados nesse trabalho foi o Método de Holt, com um RMSE de 118,06 e o método com pior acurácia foi o gerado pelo Random Forest com RMSE igual a 404,37.

REFERÊNCIAS

BREIMAN, L. ***Random Forests. Machine Learning*** (2001). Disponível em: <https://doi.org/10.1023/A:1010933404324> Acesso em Março 2020.
CORRÊA, H. L. (2014) Administração De Cadeias De Suprimento E Logística, 1ª edição, Atlas.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. ***Random Forests***. Boston, MA: Springer US, 2012. Disponível em: <https://doi.org/10.1007/978-1-4419-9326-7_5>. Acesso em Fevereiro 2021.

EHLERS, R. S. **Análise de Séries Temporais** (2007). Departamento de Estatística, UFPR. Disponível em <http://www.each.usp.br/rvicente/AnaliseDeSeriesTemporais.pdf, 2007> Acesso em Março 2020.

GARCIA, S. C. (2000) **O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde**. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/4703/000503532.pdf?sequence=1> Acesso em Fevereiro 2021.

HYNDMAN, R.J., & ATHANASOPOULOS, G. (2018) ***Forecasting: principles and practice***, 2ª edição, OTexts: Melbourne, Australia. Disponível em: <OTexts.com/fpp2> Acesso em Março 2020.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Índice Nacional de Preços ao Consumidor Amplo - IPCA. 2021**. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indicenacional-de-precos-ao-consumidor-amplo.html?=&t=conceitos-e-metodos> Acesso em Março 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Índice Nacional de Preços ao Consumidor Amplo - IPCA. 2021**. Disponível em: < https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indicenacional-de-precos-ao-consumidor-amplo.html?=&t=o-que-e> Acesso em Março 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Série Relatórios Metodológicos**. 8. ed. Rio de Janeiro, 2020. Ebook. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indicenacional-de-precos-ao-consumidor-amplo.html?=&t=notas-tecnicas> Acesso em Março 2021.

JÚNIOR, A. M. (2007) **Análise de Métodos de Previsão de Demanda Baseados em Séries Temporais em uma Empresa do Setor de Perfumes e Cosméticos.** Disponível em: <http://www.livrosgratis.com.br/download_livro_7785/analise_de_metodos_de_previsao_de_demanda_baseados_em_series_temporais_em_uma_empresa_do_setor_de_perfumes_e_cosmeticos> Acesso em Fevereiro 2021.

MONTGOMERY, D.; JENNINGS, C. ***Introduction to Time Series Analysis and Forecasting***. Hoboken, New Jersey: Wiley & Sons, Inc., 2008.

PEINADO, J.; GRAEML, A. R. (2007) Administração da Produção (Operações Industriais e de Serviços). 1ª. ed. Curitiba.

WHEELWRIGHT, S.C., MAKRIDAKIS, S. e McGee, V.E. (1983). ***Forecasting: Methods and Applications***, 2ª Ed. New York: John Wiley and Sons.