

UNIVERSIDADE FEDERAL DE MINAS GERAIS  
Instituto de Ciências Exatas  
Programa de Pós-Graduação / Especialização

Gleudson Willer Gomes

**PREVISÃO DO PREÇO DE FECHAMENTO DE UMA AÇÃO LISTADA NA BOLSA  
DE VALORES B3 (BRASIL, BOLSA, BALCÃO) COM ANÁLISE PREDITIVA E  
LINGUAGEM R**

Belo Horizonte, MG  
2022

GLEIDSON WILLER GOMES

**PREVISÃO DO PREÇO DE FECHAMENTO DE UMA AÇÃO LISTADA NA BOLSA  
DE VALORES B3 (BRASIL, BOLSA, BALCÃO) COM ANÁLISE PREDITIVA E  
LINGUAGEM R**

Monografia apresentada ao Curso de Especialização em Estatística do Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de Especialista em Estatística.

Orientadora: Profa. Dra. Ilka Afonso Reis

Belo Horizonte, MG  
Instituto de Ciências Exatas da UFMG  
2022

2022, Gleidson Willer Gomes.  
Todos os direitos reservados

Gomes, Gleidson Willer.

G633p Previsão do preço de fechamento de uma ação listada na bolsa de valores b3 (brasil, bolsa, balcão) com análise preditiva e linguagem R [manuscrito] / Gleidson Willer Gomes. — 2022.  
39.f. il.

Orientadora: Ilka Afonso Reis.  
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.  
Referências 38-39.

1. Estatística. 2. Bolsa de valores. 3. Análise de regressão. 4. Ciência de dados. 5. Random Forest. I. Reis, Ilka Afonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEX



Universidade Federal de Minas Gerais

E-mail:

Instituto de Ciências Exatas

Tel: 3409-

9-5924

Departamento de Estatística

P Programa de Pós-Graduação / Especialização

Av. Pres. Antônio Carlos, 6627 - Pampulha

31270-901 - Belo Horizonte - MG

## ATA DO 246º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE GLEIDSON WILLER GOMES.

Aos quinze dias do mês de setembro de 2022, às 14:30 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Gleidson Willer Gomes**, intitulado: “*Previsão do Preço de Fechamento de uma Ação Listada na Bolsa de Valores B3 (Brasil, Bolsa, Balcão) com Análise Preditiva e Linguagem R*”, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Ilka Afonso Reis – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 15 de setembro de 2022.

Documento assinado digitalmente



ILKA AFONSO REIS

Data: 15/09/2022 18:19:49-0300

Verifique em <https://verificador.iti.br>

Prof.<sup>a</sup> Ilka Afonso Reis (Orientadora)  
Departamento de Estatística / ICEx / UFMG

Lourdes Coral Contreras  
Montenegro

Assinado de forma digital por Lourdes  
Coral Contreras Montenegro  
Dados: 2022.09.20 11:51:58 -0300

Prof.<sup>a</sup> Lourdes Coral Montenegro  
Departamento de Estatística / ICEx / UFMG

Jussiane Nader  
Gonçalves

Assinado de forma digital por  
Jussiane Nader Gonçalves  
Dados: 2022.09.20 20:46:30 -0300

Prof.<sup>a</sup> Jussiane Nader Gonçalves  
Departamento de Estatística / ICEx / UFMG



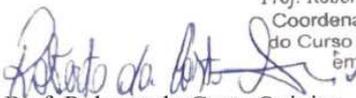
**Universidade Federal de Minas Gerais**  
**Instituto de Ciências Exatas**  
**Departamento de Estatística**  
**Programa de Pós-Graduação / Especialização**  
Av. Pres. Antônio Carlos, 6627 - Pampulha  
31270-901 – Belo Horizonte – MG

E-mail: [pgest@ufmg.br](mailto:pgest@ufmg.br)  
Tel: 3409-5923 – FAX: 3409-5924

### **DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.**

Declaro para os devidos fins que Gleidson Willer Gomes, número de registro 2020680313, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que entregou para sua orientadora, a professora Ilka Afonso Reis, que aprovou a versão final. O trabalho foi apresentado no dia 15 de setembro de 2022 com o título “Previsão do Preço de Fechamento de uma Ação Listada na Bolsa de Valores B3 (Brasil, Bolsa, Balcão) com Análise Preditiva e Linguagem R”.

Belo Horizonte, 04 de novembro de 2022

  
*Prof. Roberto da Costa Quinino*  
Coordenador da Comissão  
do Curso de Especialização  
em Estatística  
**Prof. Roberto da Costa Quinino**  
**Coordenador do curso de**  
**Especialização em Estatística**  
**Departamento de Estatística / UFMG**

## **AGRADECIMENTOS**

Agradeço à Deus pela luz e energia necessária nesta caminhada, a minha família pelo apoio incondicional e, em especial, a Professora Ilka Reis pela orientação e ensinamentos que contribuíram com o meu crescimento acadêmico. Agradeço também as Professoras Jussiane Gonçalves e Lourdes Contreras que compuseram a banca pela contribuição e comentários sobre o trabalho. Agradeço a todos os professores e colaboradores do Curso de Especialização em Estatística do Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais e a todos que contribuíram de forma direta ou indireta com a minha formação.

## RESUMO

O artigo apresenta uma análise preditiva do preço de fechamento de uma ação de uma empresa listada na B3 (Brasil, Bolsa, Balcão). A motivação foi saber o potencial de valorização ou desvalorização da ação da Itaúsa no dia seguinte. Para tal, foram comparados os modelos de Regressão Linear ajustado com o método dos Mínimos Quadrados Generalizados (MQG), *Random Forest* (MRF) e *XGBoost* (XGB) em nove pares de séries extraídas da série histórica, nos quais o primeiro elemento do par é a série de treino e o segundo, com observações consecutivas ao primeiro, é a série de teste. A métrica de desempenho utilizada foi a Raiz Quadrado do Erro Quadrático Médio e a mediana dos erros de previsão. O modelo MQG obteve o melhor desempenho entre os demais modelos. Com o modelo de predição escolhido, será possível traçar novas estratégias de investimentos em ações.

**Palavras-chave:** Mercado de Ações, Estatística, Regressão Linear, Ciência de Dados, *Random Forest*, *XGBoost*.

## ABSTRACT

The article presents a predictive analysis of the closing price of a share of a company listed on B3 (Brasil, Bolsa, Balcão). The motivation was to find out the potential for appreciation or devaluation of Itaúsa's share on the following day. For this, the Linear Regression models adjusted with the Generalized Least Squares (MQG), Random Forest (MRF) and XGBoost (XGB) method were compared in nine pairs of series extracted from the historical series, in which the first element of the pair is the training series and the second, with observations consecutive to the first, is the test series. The performance metric used was the Root Mean Squared Error and the median of the forecast errors. The MQG model obtained the best performance among the other models. With the chosen prediction model, it will be possible to outline new strategies for investing in stocks.

**Key-words:** Stock Market, Statistics, Linear Regression, Data Science, Random Forest, XGBoost.

## LISTA DE ILUSTRAÇÕES

FIGURA 1 – Segmentação do mercado financeiro .....	14
FIGURA 2 – Diagrama de uma árvore de decisão simples para diagnóstico de um paciente .....	16
FIGURA 3a – Diferença entre os métodos <i>Bagging</i> e <i>Boosting Ensemble</i> .....	18
FIGURA 3b – Evolução dos algoritmos baseados em árvores de decisão .....	18
FIGURA 4 – Simplificação do método dos MQG no método dos MQO .....	20
FIGURA 5 – Data Analytics Lifecycle .....	21
FIGURA 6 – Série histórica da ação preferencial (ITSA4) da Itaúsa no período selecionado .....	23
FIGURA 7a – Resultado para previsões do par de Séries 1 segundo modelos MQG, MRF e XGB.....	27
FIGURA 7b – Resultado para previsões do par de Séries 2 segundo modelos MQG, MRF e XGB.....	28
FIGURA 7c – Resultado para previsões do par de Séries 3 segundo modelos MQG, MRF e XGB.....	29
FIGURA 7d – Resultado para previsões do par de Séries 4 segundo modelos MQG, MRF e XGB.....	30
FIGURA 7e – Resultado para previsões do par de Séries 5 segundo modelos MQG, MRF e XGB.....	31
FIGURA 7f – Resultado para previsões do par de Séries 6 segundo modelos MQG, MRF e XGB.....	32
FIGURA 7g – Resultado para previsões do par de Séries 7 segundo modelos MQG, MRF e XGB.....	33
FIGURA 7h – Resultado para previsões do par de Séries 8 segundo modelos MQG, MRF e XGB.....	34
FIGURA 7i – Resultado para previsões do par de Séries 9 segundo modelos MQG, MRF e XGB.....	35
FIGURA 8 – Comparativo da avaliação dos modelos MQG, MRF e XGB pela métrica REQM.....	37

## LISTA DE TABELAS

TABELA 1 – Resumo da avaliação do desempenho dos modelos MQG, MRF e XGB segundo as métricas REQM e mediana dos erros de previsão em todos os pares de séries.....[36](#)

## SUMÁRIO

1.	PRÓLOGO .....	12
2.	ARTIGO: PREVISÃO DO PREÇO DE FECHAMENTO DE UMA AÇÃO LISTADA NA BOLSA DE VALORES B3 (BRASIL, BOLSA, BALCÃO) COM ANÁLISE PREDITIVA E LINGUAGEM R.....	12
2.1.	Introdução .....	12
2.2.	O Mercado Acionário.....	14
2.3.	Modelos para Previsão do Preço de Ação.....	16
2.4.	Materiais e Métodos .....	21
2.4.1.	Ciclo de Vida para Análise de Dados.....	21
2.4.2.	Estudo de Caso .....	22
2.5.	Resultados .....	25
2.6.	Conclusão.....	37
	REFERÊNCIAS.....	40

## **1. PRÓLOGO**

Este trabalho de conclusão de curso, apresentado ao curso de Especialização em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Especialista em Estatística, foi escrito no formato de manuscrito com o propósito de publicação em sites, revistas e periódicos para a área de conhecimento em Estatística e/ou Ciência da Computação. Para isso, foi utilizado o modelo para publicação de artigo da Sociedade Brasileira de Computação (SBC) pela experiência de já ter escrito um manuscrito nesse padrão. O meta-artigo pode ser consultado no seguinte endereço eletrônico: <https://www.sbc.org.br/documentos-da-sbc/summary/169-templates-para-artigos-e-capitulos-de-livros/878-modelosparapublicaodeartigos>.

## **2. ARTIGO: PREVISÃO DO PREÇO DE FECHAMENTO DE UMA AÇÃO LISTADA NA BOLSA DE VALORES B3 (BRASIL, BOLSA, BALCÃO) COM ANÁLISE PREDITIVA E LINGUAGEM R**

### **2.1. Introdução**

Em 2022, o Brasil alcançou uma marca expressiva de novos investidores no mercado de capitais. Segundo a Bolsa de Valores B3 (Brasil, Bolsa, Balcão) (2022), são 4,2 milhões de contas distintas de pessoas físicas em corretoras no Brasil atingidas no primeiro mês de 2022 com investimento em renda variável. Somados aos investidores em renda fixa, são mais de 13,1 milhões de pessoas físicas no mesmo período. Na comparação com último mês de 2020, é um crescimento de 23% de pessoas físicas na B3.

O mercado acionário é cada vez mais utilizado pelos brasileiros como instrumento de aposentadoria, evolução patrimonial e alavancagem da carteira de investimento. A democratização dos investimentos provocada pelo acesso a ativo de baixo aporte inicial, gama diversificada de produtos, cenário de juros baixos em 2020 e primeiro semestre de 2021 no Brasil contribuiu para a entrada de novos investidores na B3. A renda variável passou a fazer parte da estratégia financeira do brasileiro, que expandiu seus investimentos e ultrapassou a fronteira da tradicional caderneta de poupança. Nesse cenário, saber o preço de fechamento estimado de uma ação é identificar o potencial de valorização desse ativo no dia seguinte, ter apoio na decisão final de compra ou venda do ativo em questão por parte do investidor, que poderá ter condições de avaliar possíveis investimentos rentáveis para a sua carteira.

O objetivo deste trabalho foi comparar o desempenho dos modelos de Regressão Linear ajustado com o método dos Mínimos Quadrados Generalizados (MQG), *Random Forest* (MRF) e *XGBoost* (XGB) para prever o preço de fechamento de uma ação no dia seguinte de uma *holding* brasileira de investimentos listada na B3. Foram utilizados dados históricos do preço de fechamento (*Close*) da ação disponibilizados pela B3 ao final do dia.

Com o modelo de predição escolhido, será possível traçar novas estratégias (ou até mesmo automatizá-las), saber o potencial de valorização ou desvalorização da ação no dia seguinte e obter apoio na tomada de decisão para investimento em ação conforme a estratégia definida pelo investidor.

### **2.1.1. Motivação**

Diante do cenário apresentado pela Bolsa de Valores B3 (Brasil, Bolsa, Balcão) (2022) sobre novos investidores pessoa física no mercado de capitais, faz-se necessário o uso de uma ferramenta para apoio nas decisões de investimentos.

Com os avanços da tecnologia na área de aprendizado de máquina oriundo do advento do *Big Data* [Oracle 2022], muitos algoritmos foram adaptados para processar alto volume de dados com rapidez e eficiência.

Em consonância aos dois cenários acima, o presente trabalho utilizou a tradicional técnica de Regressão Linear ajustado com MQG em comparação com os algoritmos de aprendizado de máquina recorrentes na área de ciência de dados – o MRF e XGB, de modo a compreender o poder de precisão e confiabilidade nos resultados entre os modelos utilizados neste estudo. A motivação foi saber o potencial de valorização ou desvalorização da ação da Itaúsa no dia seguinte.

### **2.1.2. Organização do Trabalho**

O presente artigo segue organizado em mais seis seções. A Seção 2 apresenta uma visão geral sobre o mercado acionário e, em seguida, a Seção 3 aborda os modelos para previsão do preço de ação estudados neste trabalho. A Seção 4 introduz a metodologia para comparação do desempenho dos modelos, bem como os materiais e métodos utilizados. A Seção 5 descreve e discute os resultados obtidos e a Seção 6 traz as conclusões do trabalho. Finalmente, a Seção 7 propõe melhorias e trabalhos futuros.

## 2.2. O Mercado Acionário

Segundo Mercado de Valores Mobiliários Brasileiro (2019), o mercado financeiro é composto por outros quatro mercados definidos como mercado monetário, mercado de crédito, mercado de câmbio e mercado de capitais conforme a Figura 1:

**Figura 1: Segmentação do mercado financeiro.**



**Fonte: Mercado de Valores Mobiliários Brasileiro (2019).**

O mercado monetário está relacionado ao trâmite de recursos entre instituições financeiras no curto prazo e é regido pelo Banco Central.

O mercado de crédito consiste na tomada de recurso financeiro de agente superavitário para concessão de empréstimo a terceiro com a promessa de ganho de curto e médio prazo com um percentual pré-definido no ato da negociação executados por sociedades financeiras.

O mercado de câmbio contempla a negociação de moeda estrangeira e nacional entre um país e outro e é regulado pelo Banco Central do Brasil.

O mercado de capitais, tema central deste trabalho, organiza operações realizadas com ativos e produtos financeiros para captação de recursos no médio e longo prazo entre um agente superavitário e outro agente deficitário com a intermediação de uma instituição financeira que oferece toda a infraestrutura necessária para tal.

### 2.2.1. Bolsa de Valores B3 (Brasil, Bolsa, Balcão)

Segundo Portal B3 (2022), a Bolsa de Valores é a plataforma que oferece produtos financeiros para compra e venda de ativos. No Brasil, a B3 (Brasil, Bolsa, Balcão) é a empresa privada oficial que oferece toda a infraestrutura para a negociação de ativos de Renda Fixa (Títulos Privados e Públicos), Renda Variável (Ação, Fundo de Investimento Imobiliário, outros), Moedas, *Commodities* e Leilões.

Ainda segundo Portal B3 (2022), os principais produtos financeiros para investimento em renda variável são as Ações e os Fundos de Investimentos Imobiliários. As Ações são papéis emitidos por empresas com capital aberto com o objetivo de capturar recursos financeiros para investir em novos projetos com o foco no crescimento da organização, diferencial competitivo no mercado, melhoria nos processos de governança e transparência financeira. Podem ser ordinárias quando seu detentor tem a opção de participar da votação em assembleia para definir o rumo da empresa, ou preferenciais, quando se tem o direito de receber dividendos elevados em relação as ações ordinárias. Por sua vez, os Fundo de Investimento Imobiliário são constituídos por papéis emitidos destinados a aplicação no mercado imobiliário, como imóveis urbanos e rurais, títulos e cotas ligados ao setor imobiliário.

### 2.2.2. Trabalhos Relacionados no Mercado Acionário

De acordo com Ribeiro (2020), prever o preço de fechamento de uma ação é uma tarefa demandada pelos investidores. Diante disso, o autor propõe comparar oito modelos de predição para o preço de fechamento do índice Bovespa (Ibovespa): o AdaBoost (*Adaptive Boosting*), SVM (*Support Vector Machine*), LDA (*Linear Discriminant Analysis*), CART (*Classification and Regression Tree*), RF (*Random Forest*), MLP (*Multi-Layer Perceptron*), BR (*Bayesian Rigde*) e LR (*Linear Regression*). As métricas utilizadas foram o  $R^2$  (*Coefficient of Determination*), MAE (*Mean Absolute Error*) e RMSE (*Root Mean Squared Error*). Os modelos considerados mais eficientes foram o BR e LR.

Segundo Neto (2021), saber se uma ação irá se valorizar ou desvalorizar parece ser algo como prever o futuro. Baseado nisso, propõe usar os conceitos de *Machine Learning* e *Deep Learning* para prever o comportamento do preço de fechamento de uma ação na bolsa

de valores. Como resultado, o modelo de *Deep Learning* apresentou melhor desempenho em relação ao modelo de *Machine Learning*.

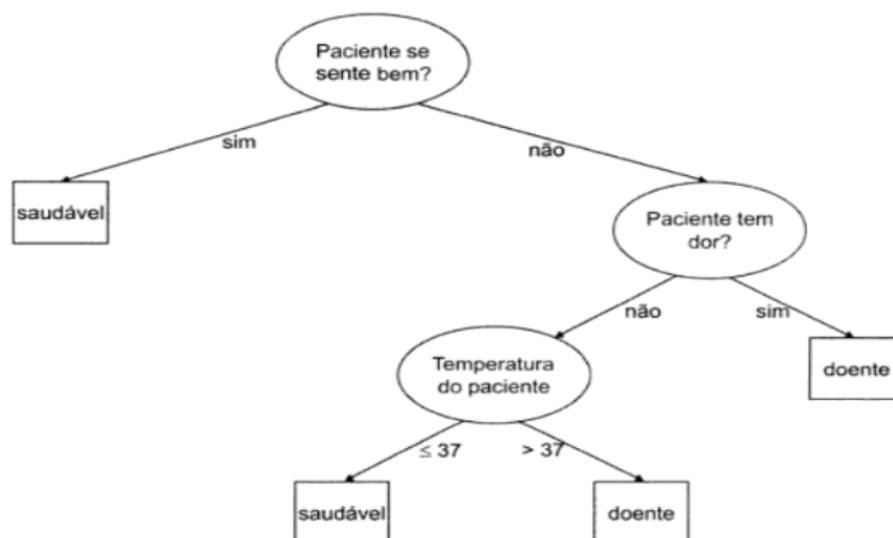
### 2.3. Modelos para Previsão do Preço de Ação

Os modelos de séries temporais são os mais utilizados para a previsão do preço de ação. O Prophet é um *framework* para previsões de séries temporais desenvolvido pelo Facebook. O Quantmod é um *framework* para modelagem financeira e quantitativa criado na linguagem R. Ambos são muito utilizados pelo mercado financeiro para prever o comportamento de uma ação no futuro.

#### 2.3.1. Árvore de Decisão, Random Forest e XGBoost

Segundo Barbieri (2011), “árvore de decisão é uma técnica que, a partir de uma massa de dados, cria e organiza regras de classificação e decisão em formato de diagramas de árvores, que vão classificar suas observações ou prever resultados futuros”. A Figura 2 ilustra a técnica para um exemplo de diagnóstico de um paciente, no qual se quer identificar se ele está saudável ou doente [Monard 2005].

**Figura 2: Diagrama de uma árvore de decisão simples para diagnóstico de um paciente.**



**Fonte: Monard (2005).**

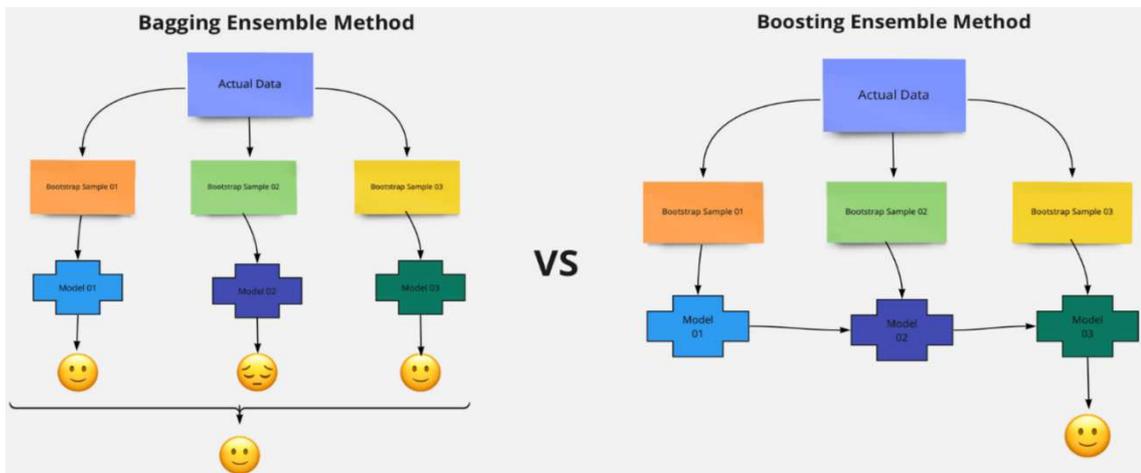
Ainda segundo Barbieri (2011), as árvores de decisão são utilizadas em diversas áreas, como na segmentação de mercado, estratificação de clientes tomadores de empréstimos em alto, médio ou baixo risco de crédito (*credit scoring*), predição de qual ação de marketing está retornando ganhos financeiros para o negócio, entre outras aplicações.

*Random Forest* ou Floresta Aleatória é um algoritmo de aprendizagem de máquina supervisionado baseado em árvore de decisão utilizado para atividades de classificação ou predição. Implementa o método *Ensemble* que consiste em combinar mais de um modelo a fim de formar um modelo mais forte. O algoritmo utiliza o método *Bagging* que gera, de forma aleatória, várias árvores de decisão com estruturas internas diferentes devido à seleção de variáveis independentes utilizadas. Cada árvore chega a uma conclusão sobre a classificação ou predição e o resultado é um consenso entre os resultados das diversas árvores geradas. [Matsumoto 2019]

*XGBoost* também é um algoritmo de aprendizagem de máquina estruturado em árvore de decisão. Implementa o método *Ensemble* a exemplo do *Random Forest* e possui um mecanismo de aumento de gradiente chamado de *Gradient Boosting*. O *Boosting* consiste em construir modelos sequenciais onde o resultado do primeiro modelo passa para o próximo com o intuito de reduzir erros e favorecer modelos de alto desempenho até que uma condição de parada seja alcançada. O aumento de gradiente consiste em minimizar erros em modelos sequenciais [XGBoost 2021]. Esse algoritmo é uma evolução dos modelos baseados em árvores de decisão, o que o torna mais competitivo e conhecido modelo no campo da ciência de dados.

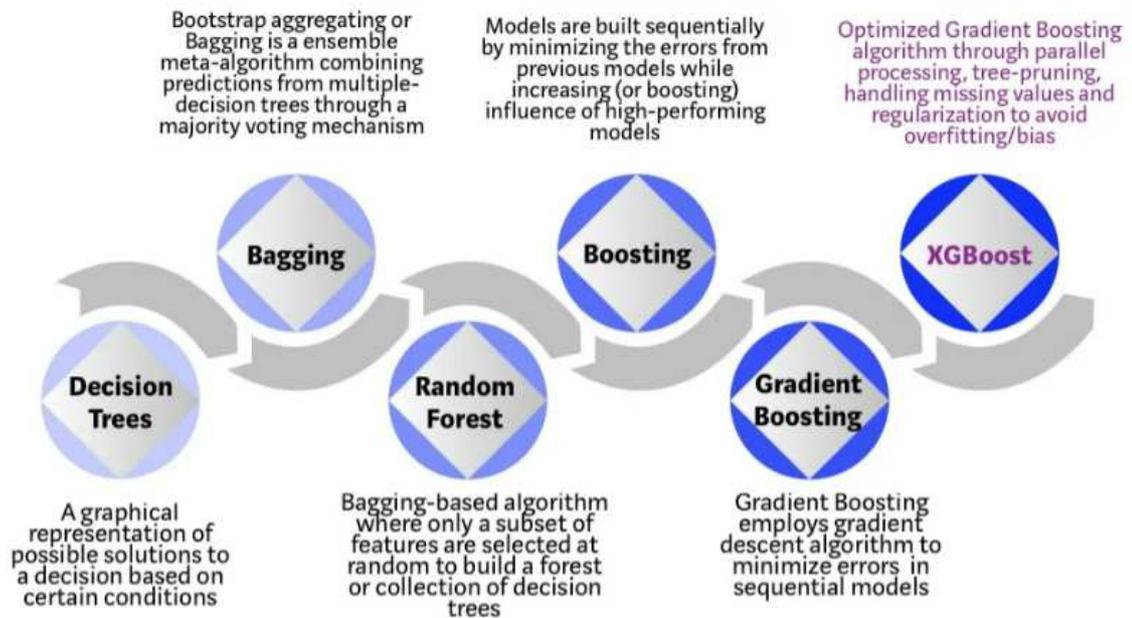
A Figura 3a ilustra a diferença entre os métodos *Bagging* e *Boosting* e a Figura 3b mostra a evolução dos algoritmos de aprendizado de máquina baseados em árvores de decisão.

**Figura 3a: Diferença entre os métodos Bagging e Boosting Ensemble.**



Fonte: Polamuri (2020).

**Figura 3b: Evolução dos algoritmos baseados em árvores de decisão.**



Fonte: Gomes (2019).

### 2.3.2. Regressão Linear

Segundo Gujarati (2011), Regressão Linear é uma técnica que permite o estudo de uma variável dependente ( $y$ ) em função de uma ou mais variáveis explicativas ( $x$ ) por meio de uma relação linear entre elas, a fim de estimar o valor médio populacional da variável  $y$  em função da(s) variável(is)  $x$  conhecidas.

O modelo de Regressão Linear Simples pode ser escrito conforme a equação abaixo:

$$Y_i = \beta_0 + \beta_1 x_i + u_i$$

Em que:

- $Y_i$  é a variável dependente;
- $\beta_0$  é o intercepto da reta (constante desconhecida a ser estimada);
- $\beta_1$  é o coeficiente angular da reta (constante desconhecida a ser estimada);
- $x_i$  é a variável explicativa;
- $u_i$  é o erro aleatório.

Como interpretação dos parâmetros do modelo linear,  $\beta_0$  representa o valor médio de  $Y_i$  quando  $x_i$  vale zero,  $\beta_1$  representa o quanto, em média,  $Y_i$  varia em função do aumento em uma unidade de  $x_i$  e  $u_i$  representa o erro aleatório para cada indivíduo  $i$  do modelo.

### **2.3.2.1. Estimação dos Parâmetros $\beta_0$ e $\beta_1$**

O ajuste do modelo de Regressão Linear Simples consiste em estimar os parâmetros  $\beta_0$  e  $\beta_1$ . Um dos métodos para isso é o de Mínimos Quadrados, no qual são encontrados os valores de  $\beta_0$  e  $\beta_1$  que minimizem a Soma dos Quadrados dos Erros (SQE).

Segundo Morettin (2017), as suposições necessárias para fazer inferências estatísticas sobre os estimadores  $\beta_0$  e  $\beta_1$  e as variáveis aleatórias envolvidas são:

- a variável  $x$  é do tipo determinística;
- para cada valor  $x$  os erros devem se distribuir ao redor da média igual a zero segundo o modelo probabilístico Normal;
- os erros  $u$  devem ter variabilidades constantes com o valor de  $x$ ;
- os erros  $u$  não devem ser correlacionados entre si.

### **2.3.2.2. O Método dos Mínimos Quadrados**

Segundo Gujarati (2011), o método dos Mínimos Quadrados Ordinários (MQO) é uma técnica de otimização matemática que procura reduzir a soma dos quadrados das diferenças

dos valores observados e estimados pelo modelo (comumente chamado de resíduos) a partir de um conjunto de dados.

Com relação às inferências a serem feitas sobre os coeficientes do modelo de regressão, o uso dos MQO pressupõe que os erros apresentem variância constante e sejam não correlacionados entre si.

Para os casos em que os erros não apresentem variância constante e/ou sejam correlacionados entre si, o método indicado é o dos MQG.

Os MQG é uma generalização dos MQO. Nos MQG a matriz de variâncias e covariâncias dos erros não necessita ter elementos iguais na diagonal (variância não constante) e os elementos fora da diagonal (covariâncias) podem ser diferentes de zero. No caso de autocorrelação temporal dos erros, os elementos fora da diagonal podem contemplar alguma estrutura de correlação no tempo, como a autorregressiva. A Figura 4 mostra as matrizes  $V$  utilizadas na composição da matriz de variâncias e covariâncias dos erros do modelo de regressão linear nos casos dos MQG e MQO:

**Figura 4: Simplificação do método dos MQG no método dos MQO.**

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} v_1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & v_2 & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & v_n \end{bmatrix}$$

Mínimos Mínimos  
Quadrados Quadrados  
Ordinários (MQO) Generalizados (MQG)

**Fonte: Adaptado de Draper (1981).**

A matriz  $V$  deve ser simétrica, assim que,  $\rho_{(i,j)} = \rho_{(j,i)}$ ,  $i = 1, 2, \dots, n$  e  $j = 1, 2, \dots, n$ . No caso de autocorrelação temporal, uma escolha bastante comum para os elementos fora da

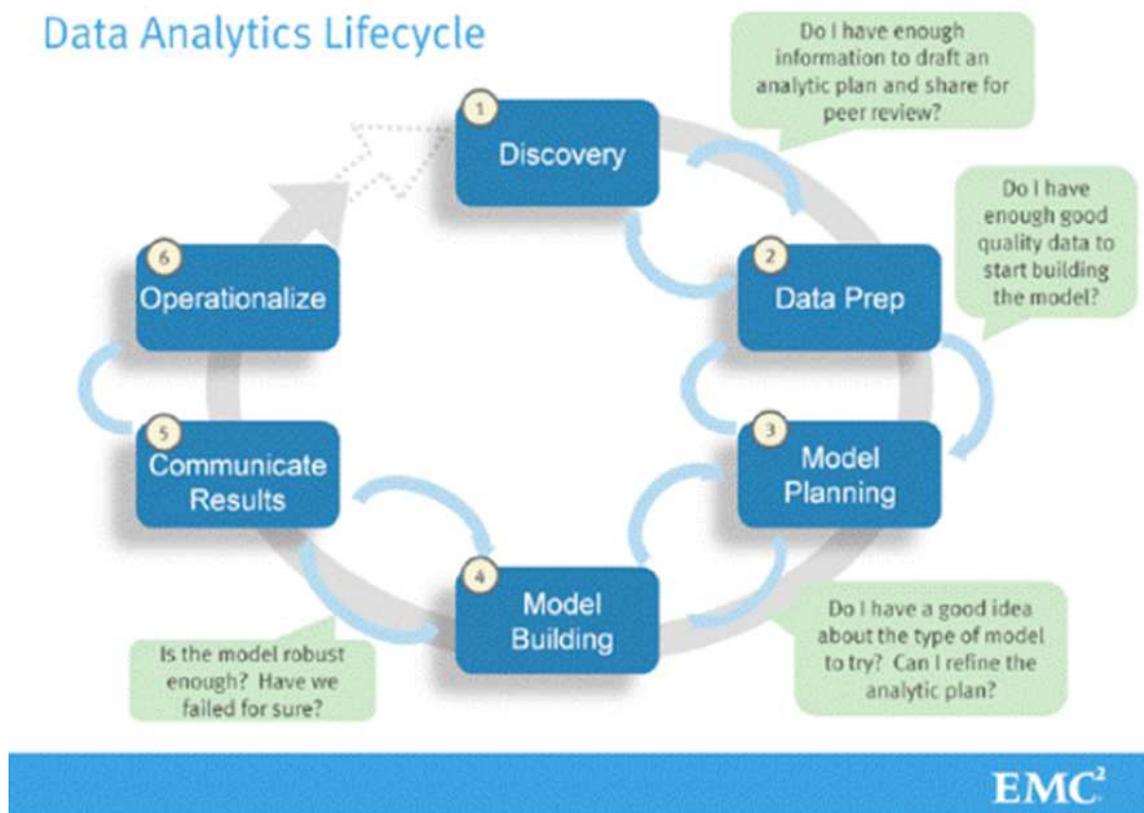
diagonal da matriz  $V$  do MQG é  $\rho_{(t,t+t)} = \rho^t, i = 1, 2, \dots, n$  e  $t = 0, 1, \dots, n - i$ , no qual  $\rho$  é o coeficiente de autocorrelação de *lag* 1, ou seja, entre  $Y_t$  e  $Y_{t+1}$ .

## 2.4. Materiais e Métodos

### 2.4.1. Ciclo de Vida para Análise de Dados

A metodologia a ser utilizada neste trabalho é o ciclo de vida para análise de dados baseado na proposta de [Todd 2012] apresentada na Figura 5.

**Figura 5: Data Analytics Lifecycle.**



**Fonte: Todd (2012).**

A primeira fase consiste em registrar a pergunta de negócio a ser respondida, levantar hipóteses iniciais para contribuir com a solução do problema e definição prévia do modelo candidato a ser utilizado no projeto analítico.

A segunda fase é o processo de preparação dos dados, onde é garantido que todas as informações necessárias são de boa qualidade, estão completas e consistentes para o desenvolvimento do modelo.

A terceira fase é o planejamento do projeto, de forma que sejam levantados potenciais modelos apropriados para resolver o problema para execução da próxima fase.

A quarta fase é a execução e avaliação do modelo, podendo ser revisto a qualquer momento durante o ciclo de vida do projeto. O objetivo principal desta fase é apurar a acurácia do modelo para o subconjunto de teste, verificar se os resultados são coerentes com o contexto do negócio para garantir que foi escolhido o modelo certo para tratar o problema de negócio.

Por fim, a quinta fase é a comunicação dos resultados alcançados e a sexta fase é a operacionalização do modelo em produção.

#### **2.4.2. Estudo de Caso**

Nesta seção estão as fases para a elaboração deste trabalho segundo o ciclo de vida para análise de dados.

Para a primeira fase, a pergunta de negócio a ser respondida é a previsão do preço de fechamento de uma ação no dia seguinte. A ação escolhida para a análise é negociada na B3 sob o código 'ITSA4' (ação preferencial) que se refere a *holding* Itaúsa que controla o banco Itaú Unibanco e demais empresas do portfólio, como a XP Inc., Alpargatas, Dexco, Aegea, Copa Energia e NTS-Nova Transportadora do Sudeste.

Para a segunda fase (preparação dos dados), os dados utilizados foram obtidos no site financeiro Yahoo! Finance [Yahoo! Finance 2022], que coleta diariamente as informações de movimentação da B3 e disponibiliza os dados históricos dos ativos negociados na bolsa de valores do Brasil. Os dados coletados para análise estão no período de 01 de abril de 2021 até 18 de agosto 2022 e abarcam 6 variáveis. As variáveis disponíveis são a data de registro das movimentações (*Date*), preço de abertura (*Open*), preço máximo (*High*), preço mínimo (*Low*), preço de fechamento (*Close*) e volume negociado (*Volume*). A Figura 6 mostra a série de dados histórica do preço de fechamento (*Close*) no período coletado disponibilizado no site Yahoo! Finance. Notamos que a série possui períodos de alta e baixa ao longo do tempo mostrando que ela possui muita volatilidade. A média do preço de fechamento (*Close*) no período é 9,93.

**Figura 6: Série histórica da ação preferencial (ITSA4) da Itaúsa no período selecionado.**



**Fonte: Yahoo! Finance (2022).**

Ainda na segunda fase, com a realização da análise exploratória dos dados, percebeu-se que as variáveis (*Open*), (*High*) e (*Low*) são muito correlacionadas com a variável (*Close*) e, portanto, não poderiam fazer parte do mesmo modelo. A variável (*Volume*) apresentou correlação baixa em relação a (*Close*) e não faz parte de nenhum modelo. Outra percepção identificada é que os dados apresentados na Figura 6 possuem correlação temporal, sendo indicado o MQG [Gujarati 2011]. Como a proposta do trabalho é prever o preço de fechamento da ação no dia seguinte, foi criado no conjunto de dados uma nova variável alvo (*Target*), que representa o preço da ação no dia seguinte em relação a (*Close*), que é o preço no dia presente. Sendo assim, as duas variáveis (*Target*) e (*Close*) são a mesma série de dados, mas defasadas por um dia uma em relação à outra. A proposta para esse trabalho foi modelar a variável (*Target*) em função da variável (*Close*), criando assim uma estrutura autorregressiva de primeira ordem (AR1) [Gujarati 2011], como mostra a seguinte equação:

$$Y_{(t)} = \beta_0 + \beta_1 * y_{(t-1)} + u_{(t)}$$

Em que:

- $Y_{(t)}$  é a série de saída (*Target*);
- $\beta_0$  é o intercepto da reta (constante desconhecida a ser estimada);
- $\beta_1$  é o coeficiente angular da reta (constante desconhecida a ser estimada);

- $y_{(t-1)}$  é a série de entrada (*Close*);
- $u_{(t)}$  é o erro aleatório do modelo.

Como interpretação dos parâmetros  $\beta_0$  e  $\beta_1$  do modelo proposto para o trabalho,  $\beta_0$  é o valor médio de  $Y_{(t)}$  quando  $y_{(t-1)}$  vale zero e  $\beta_1$  é a variação média em  $Y_{(t)}$  a cada aumento de uma unidade em  $y_{(t-1)}$ . Como se trata de uma série temporal, a autocorrelação dos erros  $u_{(t)}$  foi levada em conta ajustando o modelo de regressão por meio do MQG, no qual os elementos fora da diagonal da matriz de variâncias e covariâncias dos erros recebeu as autocorrelações temporais na forma de estrutura autorregressiva de primeira ordem (AR1). Ou seja, por exemplo, para a linha 1 da matriz V da Figura 4 tem-se  $\rho_{(1,2)} = \rho$ ,  $\rho_{(1,3)} = \rho^2$ ,  $\rho_{(1,4)} = \rho^3$ , ...,  $\rho_{(1,n-1)} = \rho^{(n-2)}$ ,  $\rho_{(1,n)} = \rho^{(n-1)}$ .

Para a terceira fase (planejamento do projeto), a proposta foi utilizar os modelos MQG, MRF e XGB como algoritmos apropriados para responder à pergunta de negócio. Devido à estrutura autorregressiva apresentada, optou-se em dividir a série da Figura 6 em nove pares de séries. O primeiro elemento do par consistiu na série de treino (com 60 dias de observações corridas) e o segundo elemento foi tomado como série de teste (com 30 dias ou mais de observações corridas e consecutivas à série de treino). O objetivo foi caracterizar as séries de treino e suas respectivas séries de teste para avaliar como o modelo treinado com uma série seria capaz de prever os valores dos 30 dias posteriores e entender o desempenho dos modelos em função da combinação dos pares de séries. A ideia foi combinar os nove pares de séries em crescente, decrescente e inexistente. A lógica para a criação dos pares foi verificar a componente de tendência dessas séries pelo método não-paramétrico de Mann-Kendall [Matos 2022], que determina se uma série possui uma tendência crescente, decrescente ou inexistente estatisticamente significativa. As hipóteses adotadas para o teste são  $H_0$  (hipótese nula): não há tendência presente nos dados (inexistente) e  $H_a$  (hipótese alternativa): há tendência presente nos dados (crescente ou decrescente). Adotou-se um nível de significância  $\alpha = 0.05$  para os testes. Para cada par de séries, foi coletado os valores da estatística de teste S (valor positivo de S indica uma tendência crescente, valor negativo de S indica uma tendência decrescente e valor zero de S indica uma tendência inexistente ao longo do tempo) e o valor-p (que é a probabilidade de se obter um valor igual ou mais extremo do que o valor de S observado na amostra, sob a hipótese nula) do método de Mann-Kendall.

Para a quarta fase (execução e avaliação do modelo), ajustou-se os modelos MQG, MRF e XGB para cada par de séries (treino e teste). A partir dos valores estimados para o preço de fechamento, calculou-se o erro de previsão (diferença entre o valor real e o valor estimado). Com os erros de previsão, foram calculadas as seguintes métricas para a avaliação

dos modelos: Raiz Quadrada do Erro Quadrático Médio (REQM) [Morettin 2017] e a mediana dos erros de previsão.

A realização da quinta fase (comunicação dos resultados) foi feita na Seção 5 deste trabalho. O ambiente utilizado para o estudo foi um computador *Desktop* com sistema operacional Windows 11, linguagem estatística R para Windows na versão 4.2.1, ambiente de desenvolvimento gráfico RStudio na versão 2022.07.1-554 para a criação das análises e demonstração dos resultados. Para execução dos pacotes na linguagem R, o pacote ‘*xlsx*’ foi usado para importar os dados para o RStudio, ‘*GGally*’ para visualizar variáveis correlacionadas, ‘*nlme*’ para execução do MQG, ‘*randomForest*’ para execução do modelo MRF e o ‘*xgboost*’ para a execução do modelo XGB. A ferramenta de *Data Discovery* Tableau Desktop 2022.1 também foi utilizada para demonstração dos resultados.

A sexta fase (operacionalização do modelo em produção) não se aplica a este trabalho de caráter científico.

## 2.5. Resultados

As figuras 7a a 7i apresentam os nove pares de séries (treino e teste), assim como sua caracterização, os resultados das previsões segundo os modelos MQG, MRF e XGB estudados, a distribuição dos erros de previsão e a comparação das séries de dados reais e previstas.

A Tabela 1 resume a avaliação dos modelos segundo as métricas REQM e mediana dos erros de previsão em todos os pares de séries.

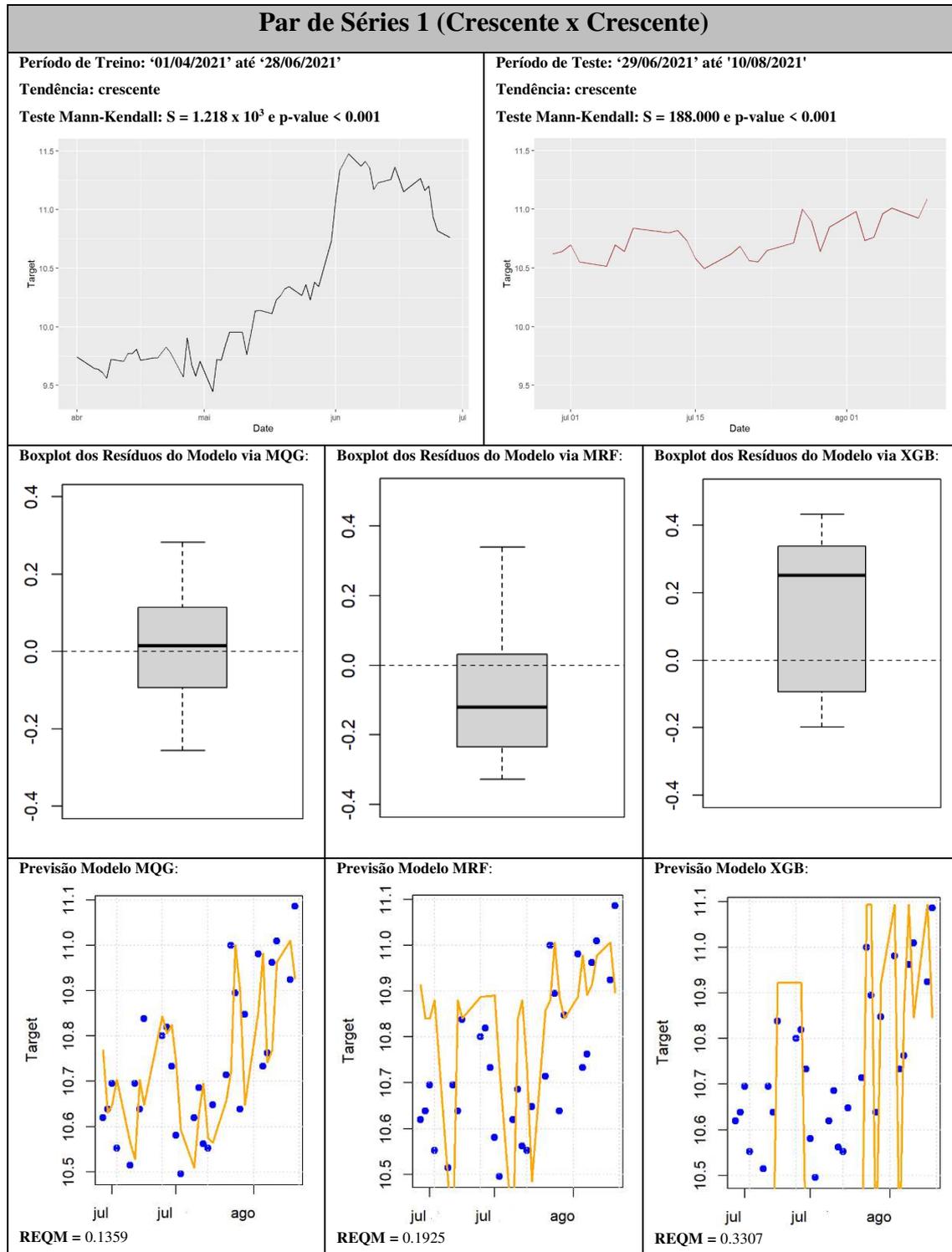
A Figura 8 mostra graficamente os resultados da comparação dos desempenhos dos modelos de previsão segundo a métrica do REQM.

Os experimentos mostram, na prática, que o MQG apresentou melhores resultados para todos os pares de séries combinados se comparado aos métodos de aprendizagem de máquina MRF e XGB para a métrica REQM. Podemos observar na Figura 8 que não há um padrão de tendência entre os pares de séries que melhore os resultados das previsões, ou seja, pares de séries com a mesma tendência (séries 1, 5 e 9) não apresentou melhor desempenho em relação aos pares séries com tendências distintas (séries 2, 3, 4, 6, 7 e 8).

Para fins de avaliação da previsão do experimento, pode-se considerar a mediana dos valores de REQM do MQG (modelo de melhor desempenho) como valor de corte aceitável para as previsões e determinar que os pares de séries com valores de REQM até 0.1866 (mediana de REQM do MQG) são considerados previsões razoáveis para o MQG, destacando

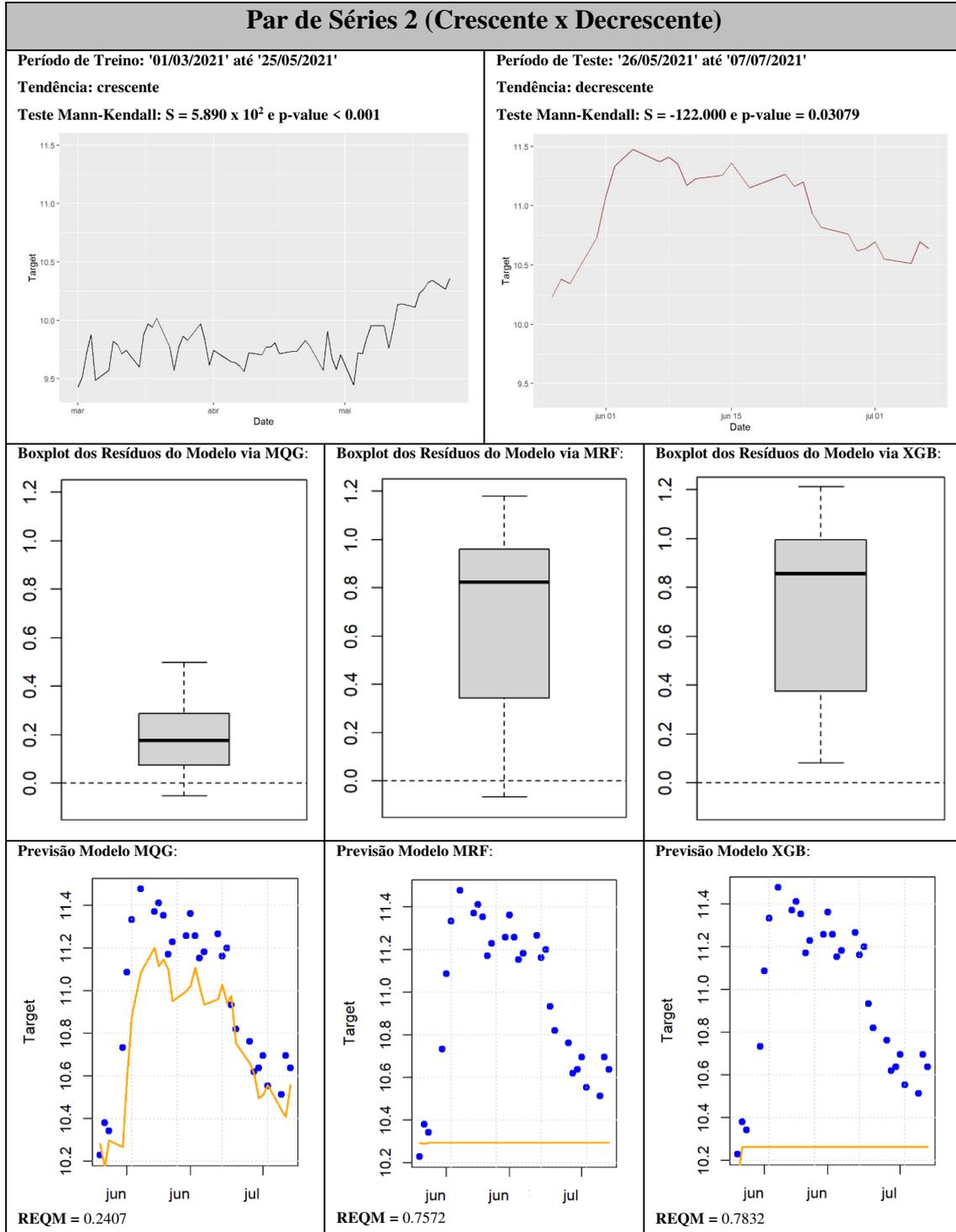
os pares de séries 1, 3, 4, 5 e 6 com suas respectivas tendências listadas na Tabela 1. Ressalta-se que fica a cargo do pesquisador definir se a previsão foi boa ou não com base na premissa de quanto ele tolera errar com o resultado alcançado.

**Figura 7a: Resultado para as previsões do par de Séries 1 segundo os modelos MQG, MRF e XGB.**



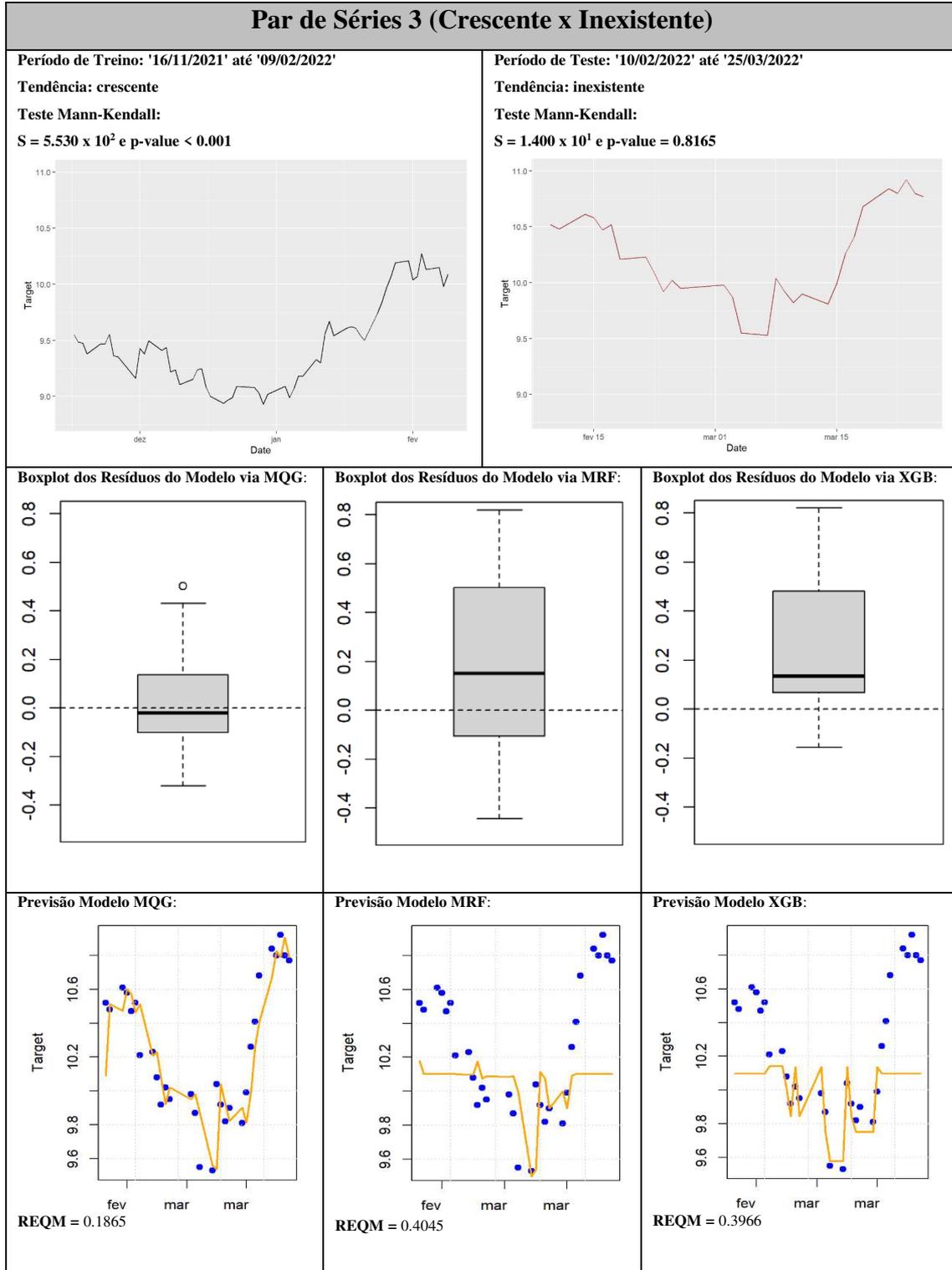
Fonte: Elaborado pelo autor (2022).

**Figura 7b: Resultado para as previsões do par de Séries 2 segundo os modelos MQG, MRF e XGB.**



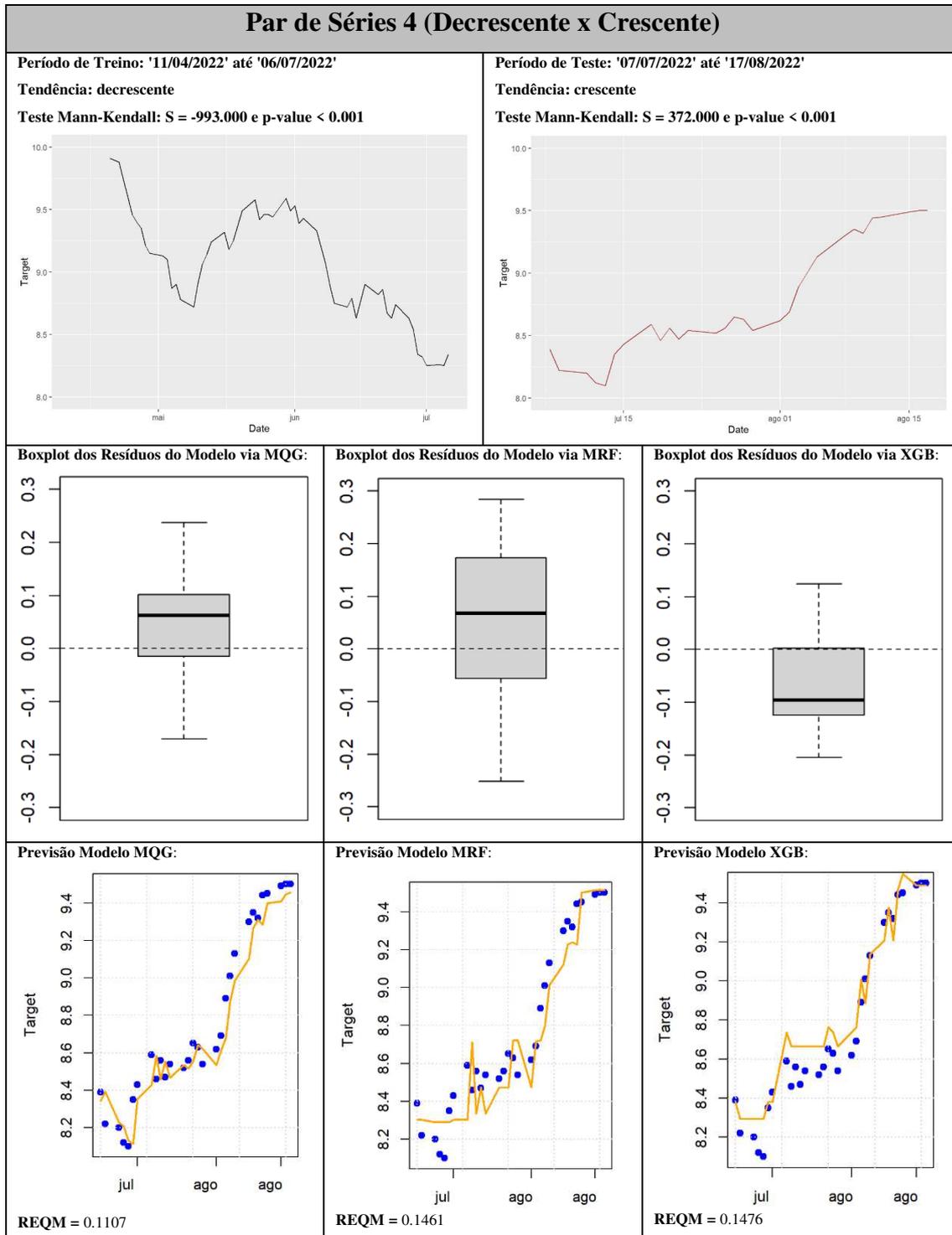
Fonte: Elaborado pelo autor (2022).

**Figura 7c: Resultado para as previsões do par de Séries 3 segundo os modelos MQG, MRF e XGB.**



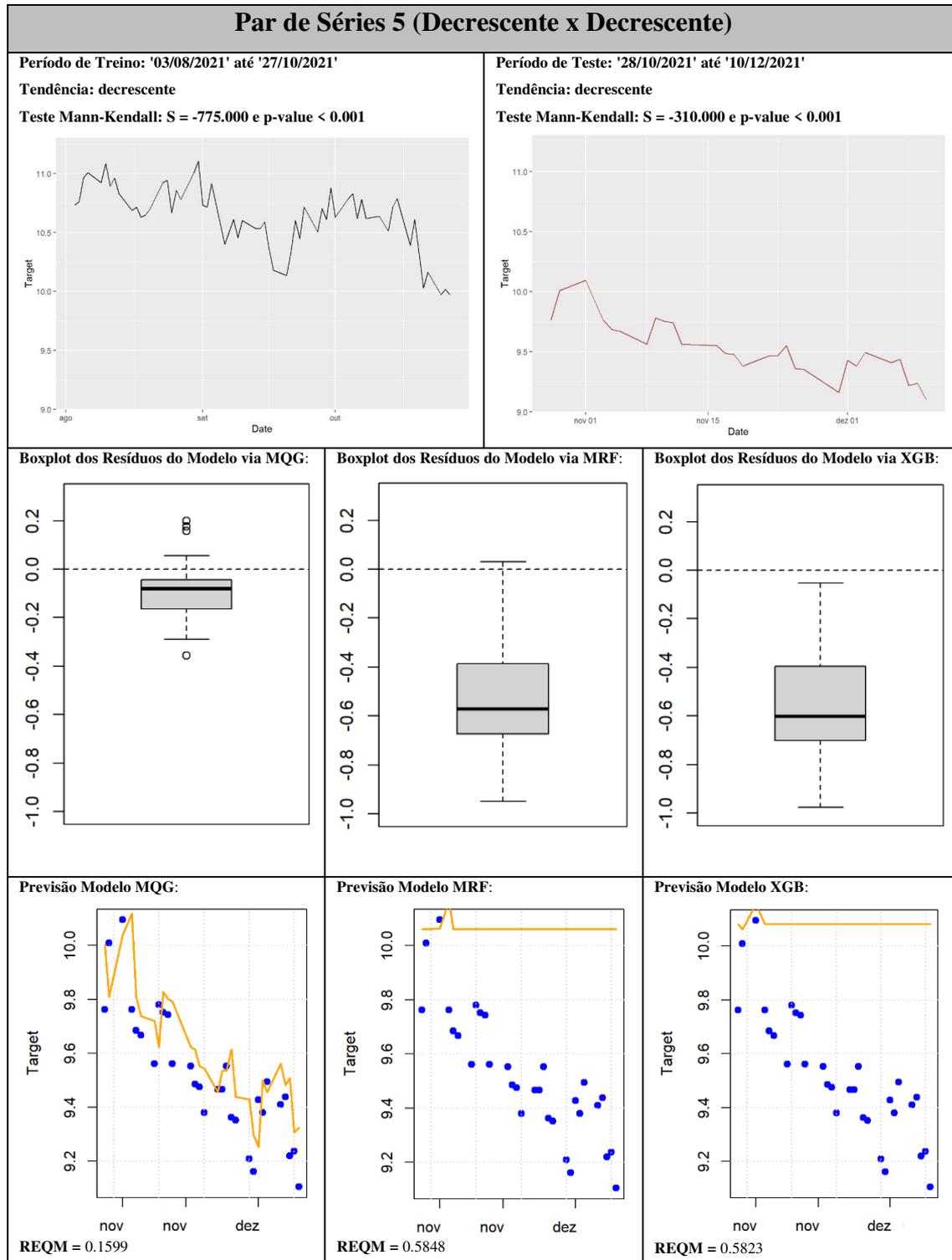
**Fonte: Elaborado pelo autor (2022).**

**Figura 7d: Resultado para as previsões do par de Séries 4 segundo os modelos MQG, MRF e XGB.**



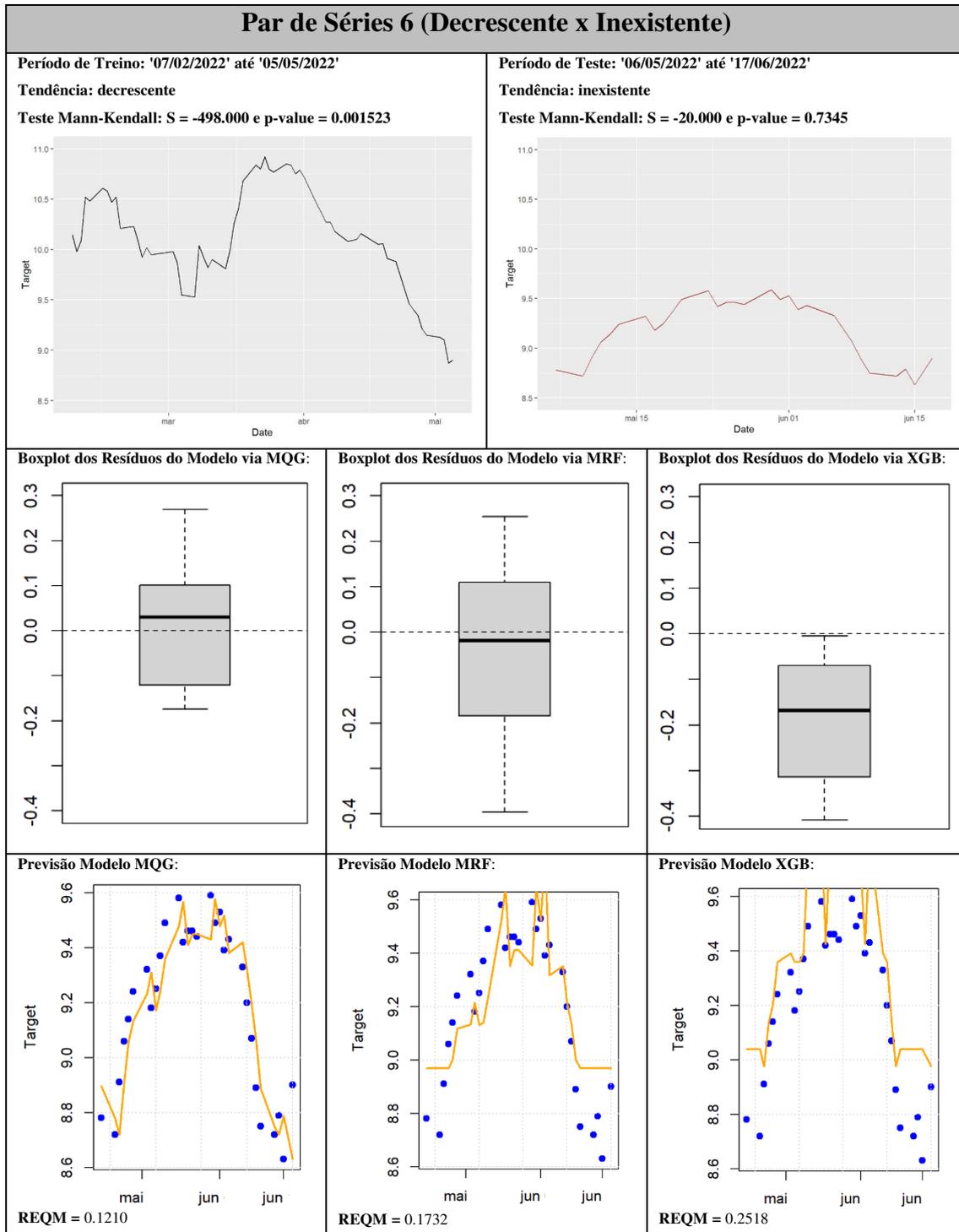
Fonte: Elaborado pelo autor (2022).

**Figura 7e: Resultado para as previsões do par de Séries 5 segundo os modelos MQG, MRF e XGB.**



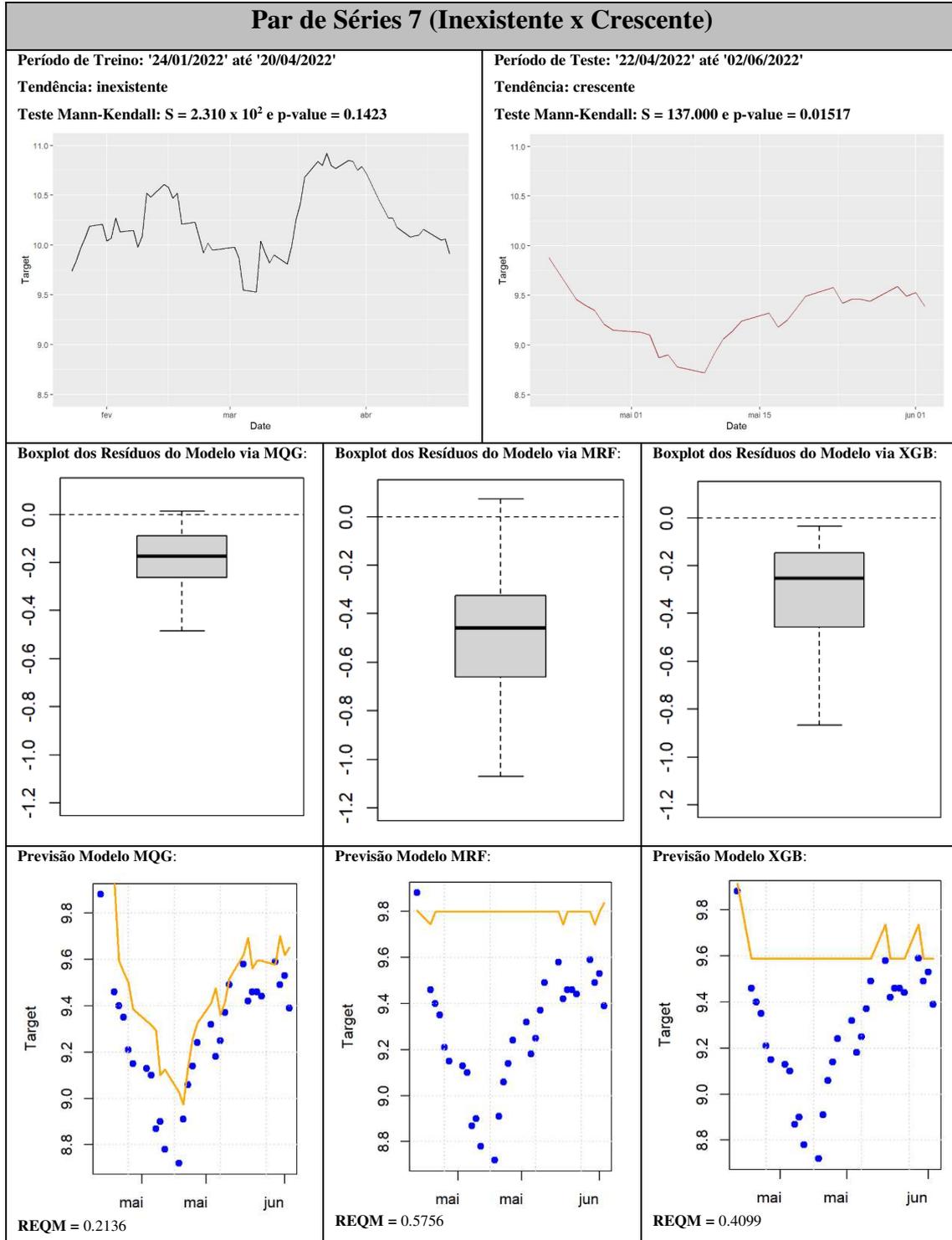
Fonte: Elaborado pelo autor (2022).

**Figura 7f: Resultado para as previsões do par de Séries 6 segundo os modelos MQG, MRF e XGB.**



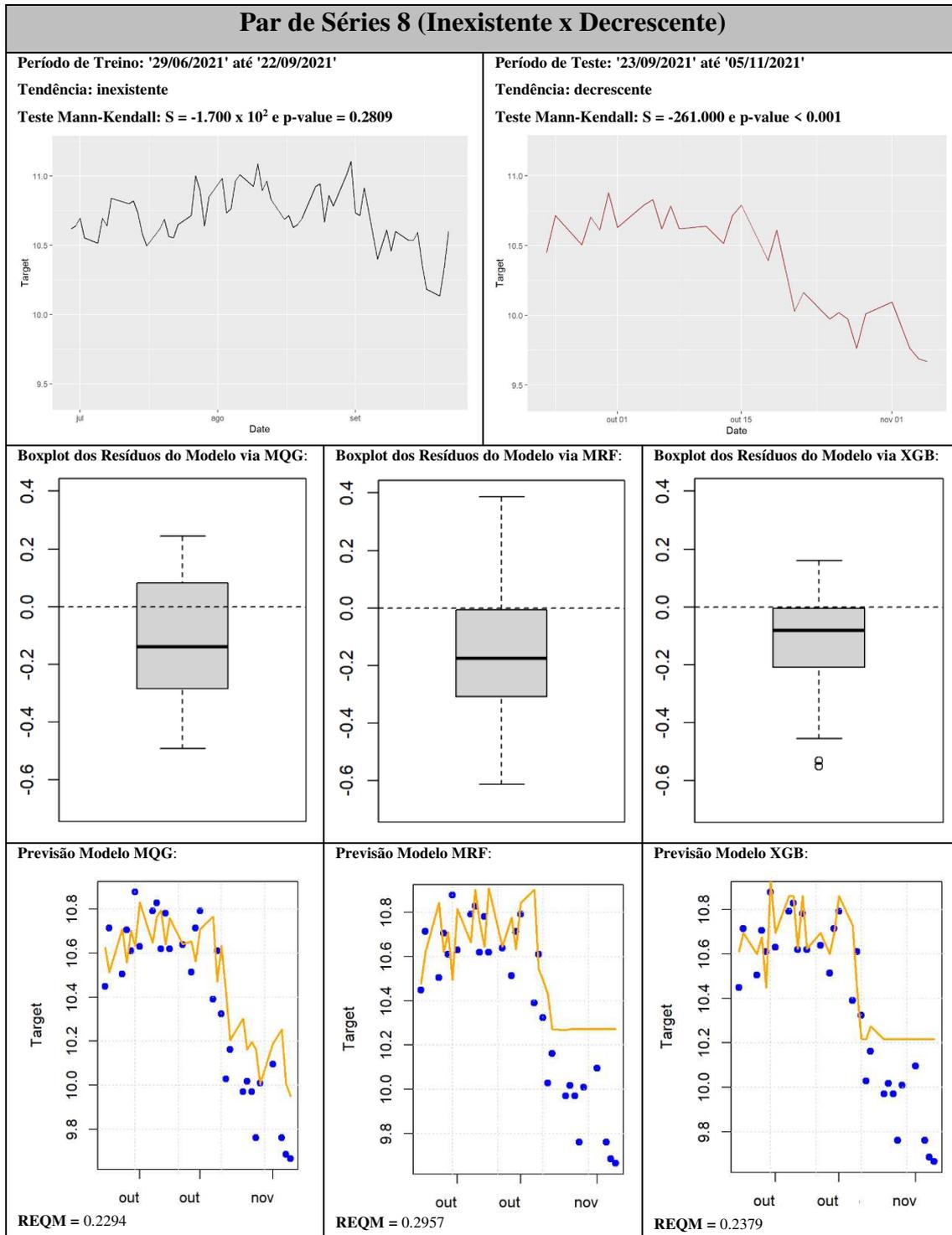
Fonte: Elaborado pelo autor (2022).

**Figura 7g: Resultado para as previsões do par de Séries 7 segundo os modelos MQG, MRF e XGB.**



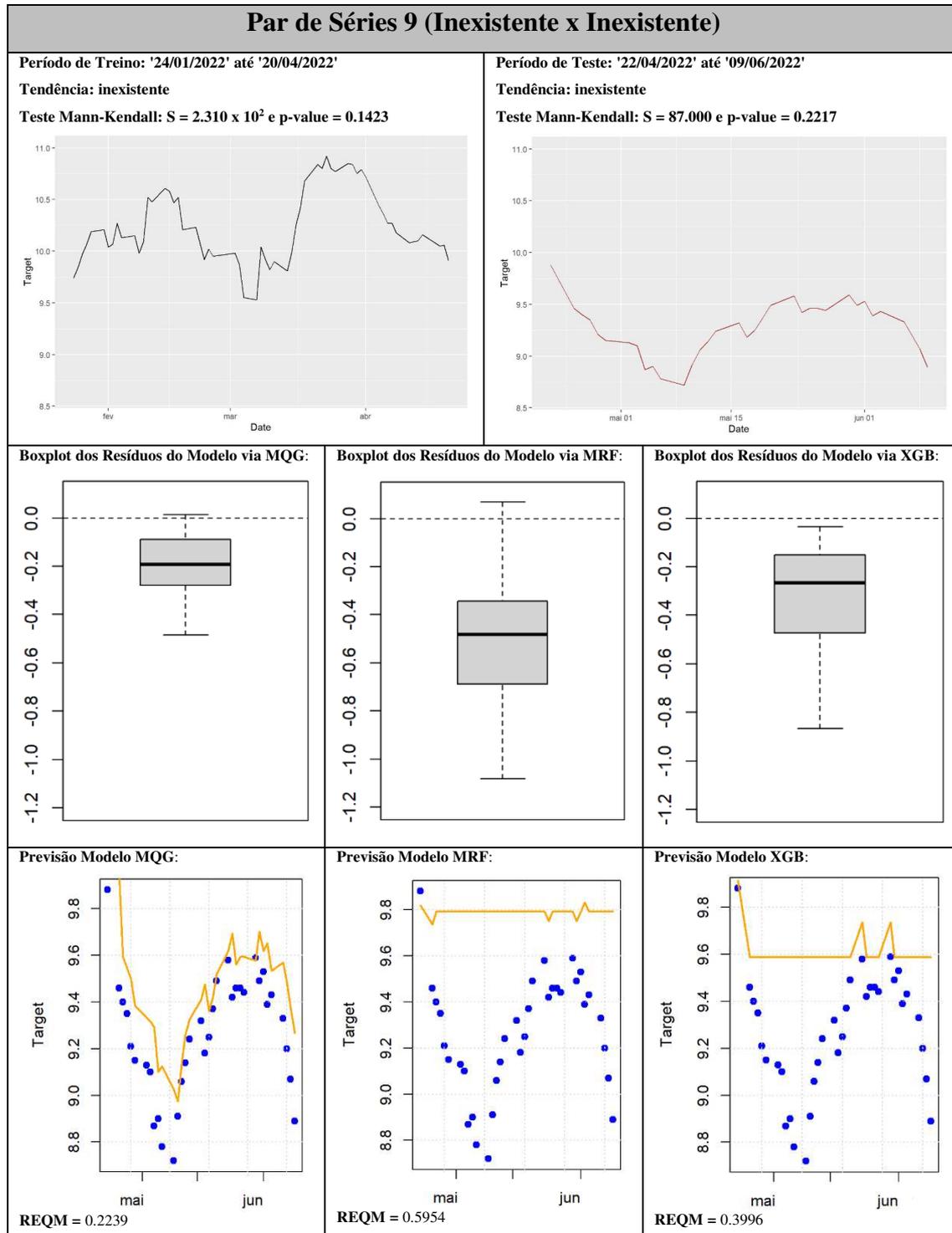
Fonte: Elaborado pelo autor (2022).

**Figura 7h: Resultado para as previsões do par de Séries 8 segundo os modelos MQG, MRF e XGB.**



Fonte: Elaborado pelo autor (2022).

**Figura 7i: Resultado para as previsões do par de Séries 9 segundo os modelos MQG, MRF e XGB (Elaborado pelo autor).**



Fonte: Elaborado pelo autor (2022).

**Tabela 1: Resumo da avaliação do desempenho dos modelos MQG, MRF e XGB segundo as métricas REQM e mediana dos erros de previsão em todos os pares de séries.**

<b>Pares de Séries</b>	<b>Caracterização / Tendência Série</b>	<b>REQM</b>	<b>Mediana Erros Previsão</b>
Série 1	Crescente x Crescente	MQG = <b>0.1359</b> MRF = 0.1926 XGB = 0.3307	MQG: 0.0107 MRF: -0.1202 XGB: 0.2513
Série 2	Crescente x Decrescente	MQG = <b>0.2407</b> MRF = 0.7552 XGB = 0.7833	MQG: 0.1770 MRF: 0.8227 XGB: 0.8564
Série 3	Crescente x Inexistente	MQG = <b>0.1866</b> MRF = 0.4064 XGB = 0.3967	MQG: -0.0202 MRF: 0.1506 XGB: 0.1354
Série 4	Decrescente x Crescente	MQG = <b>0.1108</b> MRF = 0.1472 XGB = 0.1477	MQG: 0.0626 MRF: 0.0674 XGB: -0.0955
Série 5	Decrescente x Decrescente	MQG = <b>0.1599</b> MRF = 0.5848 XGB = 0.5823	MQG: -0.0813 MRF: -0.5726 XGB: -0.6005
Série 6	Decrescente x Inexistente	MQG = <b>0.1211</b> MRF = 0.1761 XGB = 0.2518	MQG: 0.0306 MRF: -0.0186 XGB: -0.1681
Série 7	Inexistente x Crescente	MQG = <b>0.2137</b> MRF = 0.5756 XGB = 0.4100	MQG: -0.1733 MRF: -0.4593 XGB: -0.2523
Série 8	Inexistente x Decrescente	MQG = <b>0.2294</b> MRF = 0.2957 XGB = 0.2379	MQG: -0.1395 MRF: -0.1755 XGB: -0.0813
Série 9	Inexistente x Inexistente	MQG = <b>0.2239</b> MRF = 0.5954 XGB = 0.3996	MQG: -0.1934 MRF: -0.4825 XGB: -0.2673

**Fonte: Elaborado pelo autor (2022).**

**Figura 8: Comparativo da avaliação dos modelos MQG, MRF e XGB pela métrica REQM. A escala utilizada permite a comparação apenas entre linhas de uma mesma coluna.**



**Fonte: Elaborado pelo autor (2022).**

## 2.6. Conclusão

A proposta deste trabalho foi comparar os modelos MQG, MRF e XGB com a finalidade de prever o preço de fechamento da ação da Itaúsa no dia seguinte a partir da série histórica do preço dessa ação.

Para isso, foi criado no conjunto de dados uma nova variável alvo (*Target*), que representou o preço da ação no dia seguinte em relação a (*Close*), que é o preço no dia presente. Em seguida, foi modelado a variável (*Target*) em função da variável (*Close*), criando assim uma estrutura autorregressiva de primeira ordem (AR1) para o caso do modelo MQG.

Posteriormente, o desempenho da previsão dos modelos MQG, MRF e XGB foi avaliado em nove pares de séries, nos quais o primeiro elemento do par consistiu na série de treino (com 60 dias de observações corridas) e o segundo elemento foi tomado como série de teste (com 30 dias ou mais de observações corridas e consecutivas à série de treino). Cada série representa uma combinação de séries com tendências diferente, sendo crescente, decrescente e inexistente. O componente de tendência foi avaliado a partir do teste não-paramétrico de Mann-Kendall. As métricas utilizadas para a avaliação do desempenho dos modelos foram a REQM e a mediana dos erros de previsão.

Concluimos que o modelo de regressão linear com estrutura autorregressiva ajustado via MQG obteve o melhor desempenho para todas as caracterizações apresentadas nos pares de séries com resultado estatisticamente significativo se comparado aos métodos de aprendizagem de máquina MRF e XGB. Os MQG trabalham com uma estrutura na matriz de variância e covariâncias dos erros do modelo que possibilita considerar as dependências do tempo nas observações. Em contrapartida, os modelos MRF e XGB não exigem suposições a serem verificadas.

Com o modelo de predição escolhido, no caso o MQG, será possível traçar novas estratégias (ou até mesmo automatizá-las), saber o potencial de valorização ou desvalorização da ação da Itaúsa no dia seguinte e obter um apoio na tomada de decisão para investimento em ação conforme a estratégia definida pelo investidor.

Para fins de melhoria neste trabalho, recomenda-se experimentar modelos específicos que implementam o algoritmo de previsão de séries temporais. O Prophet é um modelo candidato desenvolvido e utilizado pelo Facebook. O interessante desse algoritmo é a detecção de padrões sazonais em séries temporais e não necessita de ajustes manuais, bem como a verificação de suposições [Prophet 2022]. O Quantmod é uma proposta a ser testada. É um *framework* para *Trading* criado para a linguagem R com pacote para modelagem financeira quantitativa. Possui diversas possibilidades de ajuste para uma variedade de modelos de séries temporais [Quantmod 2008]. Os modelos de séries temporais de previsão econômica, como o preço das ações, caracterizado pelo fenômeno de aglomeração de volatilidade, conhecido por grandes oscilações por longo período, são candidatos que merecem análises, como o Heterocedasticidade Condicional Autorregressiva (ARCH) e Heterocedasticidade Condicional Autorregressiva Generalizada (GARCH), que podem ter bom desempenho na identificação da volatilidade dos preços [Gujarati 2011].

Outra melhoria seria criar pares de séries novas de forma a aumentar o número de pares de séries para não depender de apenas um par de séries de cada tipo.

Finalmente, pode-se explorar outras empresas do mesmo segmento e aplicar esta metodologia para a previsão do preço de fechamento de uma ação no dia seguinte, a fim de ampliar os experimentos aqui realizados e descobrir novas aplicabilidades para os modelos estudados.

## REFERÊNCIAS

- Barbieri, C. (2011) “BI2 - Business Intelligence: Modelagem & Qualidade”. Rio de Janeiro: Elsevier. P141-145.
- Bolsa de Valores B3 (Brasil, Bolsa, Balcão), (2022) “B3 atinge 5 milhões de contas de investidores em renda variável em janeiro”. [https://www.b3.com.br/pt\\_br/noticias/5-milhoes-de-contas-de-investidores.htm](https://www.b3.com.br/pt_br/noticias/5-milhoes-de-contas-de-investidores.htm), Acesso em abril de 2022.
- Draper, N. e Smith, H. (1981) “Applied Regression Analysis”. New York: John Wiley. P221-224.
- Gomes, P. C. T. (2019) “Conheça o algoritmo XGBoost”. <https://www.datageeks.com.br/xgboost/>, Acesso em agosto de 2022.
- Gujarati, Damodar N. (2011) “Econometria Básica”. Porto Alegre: AMGH. P39, P78-83, P421-424, P435-439, P441-448, P784-790.
- Matos, Sérgio (2022) “Mann-Kendall: o off-road dos métodos de análise de tendência”. <https://www.pangeo.com.br/single-post/mann-kendall-o-off-road-dos-m%C3%A9todos-de-an%C3%A1lise-de-tend%C3%Aancia>, Acesso em agosto de 2022.
- Matsumoto, F. (2019) “Modelos de Predição | Random Forest”, <https://medium.com/turing-talks/turing-talks-18-modelos-de-predi%C3%A7%C3%A3o-random-forest-cfc91cd8e524>, Acesso em agosto de 2022.
- Mercado de Valores Mobiliários Brasileiro, (2019) “Mercado de Valores Mobiliários Brasileiro”. Rio de Janeiro: Comissão de Valores Mobiliários. P27-27, P30-33.
- Monard, M. C.; Baranauskas, J. A. (2005) “Sistemas Inteligentes - Fundamentos e Aplicações”. Barueri-SP: Manole Ltda. P120.
- Morettin, Pedro A. (2017) “Estatística Básica”. São Paulo: Saraiva. P462-467, P310-315.
- Neto, D. F. G. (2021) “Previendo valores de ações usando Machine Learning e Deep Learning”. <https://danielofelipeneto.medium.com/previendo-valores-de-a%C3%A7%C3%B5es-usando-machine-learning-e-deep-learning-80216e55e582>, Acesso em setembro de 2022.
- Oracle (2022) “O que é Big Data?”. <https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em setembro de 2022.
- Polamuri, S. (2020) “Ensemble methods | Bagging vs Boosting difference”. <https://dataaspirant.com/ensemble-methods-bagging-vs-boosting-difference/>, Acesso em setembro de 2022.
- Portal B3, (2022) “Produtos e Serviços”. [https://www.b3.com.br/pt\\_br/](https://www.b3.com.br/pt_br/), Acesso em junho de 2022.
- Prophet, (2022) “Forecasting at scale”. <https://facebook.github.io/prophet/>, Acesso em setembro de 2022.

Quantmod, (2008) “Quantitative Financial Modelling & Trading Framework for R”.  
<http://www.quantmod.com/>, Acesso em setembro de 2022.

Ribeiro, L. G. (2020) “Preveno o fechamento da bolsa de valores com Machine Learning”.  
<https://pt.linkedin.com/pulse/preveno-o-fechamento-da-bolsa-de-valores-com-machine-ribeiro>, Acesso em setembro de 2022.

Todd, S. (2012) “Data Science and Innovation”.  
[http://stevetodd.typepad.com/my\\_weblog/2012/10/data-science-and-innovation.html](http://stevetodd.typepad.com/my_weblog/2012/10/data-science-and-innovation.html), Acesso em junho de 2022.

XGBoost, (2021) “XGBoost Documentation”. <https://xgboost.readthedocs.io/en/stable/>, Acesso em agosto de 2022.

Yahoo! Finance, (2022) “Yahoo! Finance”. <https://finance.yahoo.com/>, Acesso em julho de 2022.