

UNIVERSIDADE FEDERAL DE SANTA CATARINA

**ANÁLISE COMPARATIVA DE TÉCNICAS PARA A PREVISÃO DE
SÉRIES TEMPORAIS NO CONTEXTO DE MERCADOS FINANCEIROS**

Bruno Aurélio Rôzza de Moura Campos

Florianópolis - SC

2020/2

UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE SISTEMAS DE INFORMAÇÃO

ANÁLISE COMPARATIVA DE TÉCNICAS PARA A PREVISÃO DE SÉRIES
TEMPORAIS NO CONTEXTO DE MERCADOS FINANCEIROS

Bruno Aurélio Rôzza de Moura Campos

Trabalho de conclusão de curso apresentado como
parte dos requisitos para obtenção do grau de em
Bacharel em Sistemas de Informação.

Florianópolis

2020/2

Bruno Aurélio Rôzza de Moura Campos

ANÁLISE COMPARATIVA DE TÉCNICAS PARA A PREVISÃO DE SÉRIES
TEMPORAIS NO CONTEXTO DE MERCADOS FINANCEIROS

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de em Bacharel em Sistemas de Informação.

Orientador:

Prof. Dr. Élder Rizzon Santos

Banca Examinadora:

Prof. Dra. Andrea Cristina Konrath

Prof. Dr. Guilherme Alex Derenievicz

Agradecimentos

Primeiramente agradeço à minha companheira Malu por sempre acreditar em mim. Também agradeço ao meu orientador Élder Rizzon Santos por aceitar conduzir o meu trabalho de conclusão de curso.

"Você, eu, ninguém vai bater tão duro como a vida mas não se trata de bater duro, se trata de quanto você aguenta apanhar e seguir em frente, o quanto você é capaz aguentar e continuar tentando, é assim que se consegue vencer."

Rocky Balboa

RESUMO

A previsão de Séries Temporais é um importante campo de estudo em Aprendizado de Máquina e Aprendizado Profundo. A literatura mostra que para tratar este problema foram utilizadas primeiramente técnicas vindas da econometria como os modelos ARIMA e suas derivações. Com a evolução computacional e os avanços das técnicas de aprendizagem nasceu uma nova abordagem para prever Séries Temporais. A partir disso, é possível utilizar diversas técnicas para realizar as previsões e comparar dentre elas qual tem os melhores resultados em cada contexto.

Neste trabalho o foco foi em mercados financeiros com Séries Temporais estocásticas. Foi analisada e feita previsões a partir do histórico de variação de preço da empresa de capital aberto com maior participação relativa no índice Bovespa, a Vale (VALE3).

Tendo como objetivo a comparação das principais técnicas de previsão para Séries Temporais no contexto de mercado financeiro foram realizadas análises qualitativas para compreender o estado da arte sobre previsão de Séries Temporais e teorias de previsão nos mercados financeiros. Além disso, foram realizados processos de obtenção, preparação e modelagem para garantir uma padronização dos dados de entrada em cada modelo utilizado. Por fim, foi realizada uma análise comparativa dos resultados dos preditores.

Palavras-chave: Séries Temporais, Aprendizado de Máquina, Aprendizado Profundo, ARIMA, Floresta Aleatória, SVM, LSTM.

ABSTRACT

Time Series forecasting is an important field of study in Machine Learning and Deep Learning. The literature shows that to address this problem, techniques from econometrics such as ARIMA models and their derivations were used first. With the computational evolution and the advances in learning techniques, a new approach to predict Time Series was born. From this, it is possible to use several techniques to make the predictions and compare among them which has the best results in each context.

In this work the focus was on financial markets with stochastic Time Series. Forecasts were analyzed and made based on the history of price variation of the publicly traded company with the largest relative share in the Bovespa index, Vale (VALE3).

In order to compare the main prediction techniques for Time Series in the financial market context, qualitative analysis were carried out to understand the state of the art on Time Series prediction and forecasting theories in the financial markets. In addition, processes for obtaining, preparing and modeling were carried out to ensure a standardization of the input data in each model used. Finally, a comparative analysis of the predictors' results was performed.

Keywords: Time Series, Machine Learning, Deep Learning, ARIMA, Random Forest, SVM, LSTM.

LISTA DE FIGURAS

Figura 01 - Pesquisas no Google por Ibovespa	7
Figura 02 - ST do papel VALE3	19
Figura 03 - Decomposição aditiva da ST do papel VALE3	21
Figura 04 - Decomposição multiplicativa da ST do papel VALE3	22
Figura 05 - Hiperplano de separação ótima e seus hiperplanos de suporte	26
Figura 06 - Hierarquia de abordagens para predição de ST	33
Figura 07 - Processo de previsão proposto pelo autor	39
Figura 08 - Ciclo iterativo de Box-Jenkins	44
Figura 09 - Etapas para realizar a previsão de ST	45
Figura 10 - Gráfico de distribuição do preço de fechamento das ações da VALE3 em um período de 10 anos	50
Figura 11 - Decomposição da ST	51
Figura 12 - Autocorrelação do preço de fechamento com um atraso de 313 dias.....	52
Figura 13 - Autocorrelação do preço de fechamento da VALE3	53
Figura 14 - Distribuição dos dados de treinamento	54
Figura 15 - Distribuição dos dados de teste	54
Figura 16 - Comparação entre dados de testes e dados previstos pelo modelo AR	54
Figura 17 - Comparação entre dados de testes e dados previstos pelo modelo ARIMA	55
Figura 18 - Comparação entre dados de testes e dados previstos pelo modelo SARIMA	56
Figura 19 - Comparação entre dados de testes e dados previstos pelo modelo de Floresta Aleatória	58
Figura 20 - Comparação entre dados de testes e dados previstos pelo modelo SVR	59
Figura 21 - Comparação entre dados de testes e dados previstos pelo modelo de LSTM	60
Figura 22 - Convergência da taxa de perda entre os dados de treinamento e validação	61

LISTA DE TABELAS

Tabela 01 - As cinco empresas com maior participação no índice Bovespa durante o período de maio à agosto de 2020	16
Tabela 02 - Resultados da pesquisa bibliográfica	29
Tabela 03 - Comparação dos métodos de predição	34
Tabela 04 - Resultados dos métodos de predição por cada ação	36
Tabela 05 - Análise dos métodos, métricas e fontes de dados	39
Tabela 06 - Amostra dos dados da VALE3, obtidos através do Metatrader	44
Tabela 07 - Medidas de posição dos dados do preço de fechamento das ações da VALE3	46
Tabela 08 - Resultado da autocorrelação dos dados do preço de fechamento das ações da VALE3	49
Tabela 09 - Comparação dos resultados de MAPE e RSE de cada método de predição	60

LISTA DE ABREVIATURAS E SIGLAS

ST.....	Séries Temporais
IA	Inteligência Artificial
IBOVESPA.....	Índice Bovespa
ARIMA	Autorregressivo Integrado de Médias Móveis
SARIMA	Autorregressivo Integrado de Médias Móveis Sazonal
SVM	Máquinas de Suporte Vetorial
SVR	Máquinas de Suporte Vetorial Regressiva
CVM	Comissão de Valores Mobiliários
HMA.....	Hipótese de Mercados Adaptativos
HEM	Hipótese de Eficiência de Mercados
B3.....	Brasil, Bolsa, Balcão
AP.....	Aprendizado Profundo
RNA.....	Redes Neurais Artificiais
RNN.....	Redes Neurais Recorrentes
MACD.....	Moving Average Convergence Divergence
RSI.....	Índice de Força Relativa
CCI.....	Commodity Channel Index

SUMÁRIO

1 INTRODUÇÃO	12
1.1 OBJETIVOS	14
1.1.1 Geral	15
1.1.2 Específicos	15
1.2 ESCOPO DO TRABALHO	15
1.3 MÉTODO DE PESQUISA	16
1.4 ORGANIZAÇÃO DO TRABALHO	17
2 FUNDAMENTAÇÃO TEÓRICA	18
2.1 Séries Temporais	18
2.1.1 Componentes de Séries Temporais	19
2.1.2 Decomposição	19
2.2 PREVISORES ESTATÍSTICOS	22
2.3 PREVISORES DE APRENDIZADO DE MÁQUINA	23
2.3.1 Máquina de Suporte de Vetores	23
2.4 PREVISORES DE APRENDIZADO PROFUNDO	25
2.4.1 Uso de RNA para a previsão de ST	25
2.5 AVALIAÇÃO DOS RESULTADOS DE PREVISÃO	26
3 TRABALHOS RELACIONADOS	28
3.1 Revisão Sistemática	28
3.2 Predição de Séries Temporais por Similaridade	30
3.2 Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques	32
3.3 Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques	35
3.4 Trabalhos Complementares com ênfase em Métodos Econométricos	39
3.5 Trabalhos Complementares com ênfase em AM e AP	40
3.6 Considerações Finais	40
4 DESENVOLVIMENTO	43
4.1 Obtenção, Manipulação e Pré-processamento dos Dados	45
4.2 Análise Exploratória da ST	46
4.2.1 Resumo dos Dados	47
4.2.2 Características da Série Temporal	47
4.2.3 Autocorrelação	49

4.3 Divisão e Treinamento dos Dados	51
4.4 Implementação dos Previsores Estatísticos	53
4.4.1 Modelo AR	53
4.4.1.1 Análise dos Resultados	54
4.4.2 Modelo ARIMA	54
4.4.2.1 Análise dos Resultados	54
4.4.3 SARIMA	55
4.4.3.1 Análise dos Resultados	55
4.5 Implementação dos Previsores de AM	56
4.5.1 Floresta Aleatória	56
4.5.1.1 Análise dos Resultados	57
4.5.2 SVR	57
4.5.2.1 Análise dos Resultados	58
4.6 Implementação dos Previsores de AP	58
4.6.1 Análise dos Resultados	59
4.7 Considerações Finais	60
5 CONCLUSÕES E ATIVIDADES FUTURAS	62
REFERÊNCIAS	64
APÊNDICE A: Artigo	67

1 INTRODUÇÃO

Prever movimentações nos mercados financeiros é uma tarefa complicada por causa das incertezas envolvidas. Segundo Patel et al, (2015), existem dois tipos de análises que os investidores podem realizar antes de investir em uma determinada ação. O primeiro tipo é a análise fundamentalista. Nesta modalidade, os investidores analisam o valor intrínseco das ações, levando em conta fatores como, o desempenho de produção, as vendas, a saúde econômica da empresa e clima político entre outros pontos para decidir sobre a compra ou venda do ativo. Outro tipo de análise é a técnica. Neste caso, a avaliação das ações leva em conta o estudo da variação do preço e volume ao longo do tempo.

Nos últimos três anos, a análise de ações e índices têm elevado o interesse de cada vez mais pessoas, como pesquisadores e especuladores. Este fato pode ser evidenciado no contexto brasileiro, pela quantidade de pesquisas realizadas no site da Google conforme mostrado na figura 01, com a descrição “IBOVESPA”, o principal índice brasileiro.



Figura 01. Pesquisas¹ no Google por Ibovespa².

Conforme definido por BOX et al, (2015), as Séries Temporais (ST) são uma sequência de observações ao longo do tempo. Um exemplo de ST, são as ST financeiras o qual representam

¹ Eixo y: Os números representam o interesse de pesquisa em relação ao ponto mais alto no gráfico para determinada região e hora. Um valor de 100 é o pico de popularidade do termo.

² Origem de dados: Google Trends. Disponível em www.google.com/trends.

os dados de variação do preço das ações ao decorrer de períodos. É por essa razão que o estudo delas se torna tão relevante nas estratégias de negociação.

Contudo, prever valores resultantes no tempo não é uma tarefa simples, ainda mais no contexto de mercados financeiros. A razão se deve às ST financeiras que por natureza são estocásticas, ruidosas, e não estacionárias como explicado por [Manav Kaushik et al, \(2020\)](#). Estas características corroboram a Hipótese de Mercados Adaptativos (HMA) proposta por [Lo \(2004\)](#), que trata da conciliação da Hipótese de Eficiência de Mercados (HEM) com a economia comportamental, onde a oscilação de valor de uma ação pode ser afetada não somente pelo comportamento do histórico do preço mas também por informações macroeconômicas e sentimentos de investidores, tais como aversão à perda e excesso de confiança. Sabendo disso, é possível observar ao longo do tempo que determinados padrões de tomada de decisão dos investidores tornam-se recorrentes, isso faz com que sejam criadas várias estratégias para maximizar os lucros em negociações.

Segundo [Lo \(2004\)](#), como nem toda informação financeira é explícita no mercado, ineficiências acabam afetando as estratégias de negociação. É neste momento que o uso de métodos computacionais pode se tornar uma vantagem competitiva. O mercado financeiro é constantemente estudado utilizando a Inteligência Artificial (IA) junto com a estatística.

A IA pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente ([Luger, 2013](#)). Dentro da IA existe o campo de estudos de Aprendizado de Máquina (AM) que faz uso de algoritmos e métodos que possibilitam computadores a aprender por experiências e dados analíticos ([Luger, 2013](#)).

Em cenários onde o volume, velocidade e variedade dos dados são fatores determinantes para o lucro, se torna inviável fazer análises somente com indicadores estatísticos, é então neste momento que se faz necessário o uso do AM. Além desse motivo, há fatores do comportamento humano, conforme mencionado na HMA que afetam a variação de preço. Por estes aspectos, utilizar meios de processar, reconhecer padrões e testar estratégias pode ser fundamental para negociações de sucesso.

O AM e o aprendizado profundo (AP) são subáreas da Inteligência Artificial que oferecem suporte para resolver problemas como a previsão de ST financeiras. Devido a isto,

várias técnicas de AM vêm sendo testadas para melhor prever as ST como por exemplo, máquinas de vetor de suporte, floresta aleatória, redes bayesianas e até técnicas de AP, como Redes Neurais Artificiais (Patel et al, 2015).

Embora mais antigos, modelos estatísticos comumente utilizados na área econométrica, ainda são estudados e utilizados com resultados satisfatórios para prever ST. Entre os principais preditores, podemos citar modelos lineares como AR (Autoregressive), MA (Moving Average) e ARIMA (Autoregressive Integrated Moving Average) e não-lineares como GARCH (Generalized ARCH). Partindo disso, as técnicas econométricas tradicionais, também chamadas de técnicas estatísticas, estabelecem referências econômicas devido à sua ampla utilização na literatura (Manav Kaushik et al, 2020). Mesmo assim, cabe ressaltar que estes modelos apresentam limitações devido a duas hipóteses que sempre os acompanham: (i) os dados a serem preditos são lineares e (ii) a ST é estacionária (Cao e Tay, 2001).

Segundo Patel et al, (2015), cada algoritmo tem suas próprias características para resolver o problema de predição de ST ao mesmo tempo que tem suas limitações. Por consequência destas propriedades, se torna fundamental para efeito comparativo que os dados utilizados sejam preparados de forma homogênea a todos os preditores. Fazer a preparação dos dados como, por exemplo, identificar recursos que realmente têm importância, remover valores sem peso na tomada de decisão e testar os resultados com os mesmos critérios é fundamental para garantir resultados mais precisos.

Este trabalho visa analisar as principais modelagens matemáticas para a previsão de ST financeiras. Serão realizadas etapas de exploração e preparação dos dados para que depois sejam analisados os algoritmos mais adequados a serem implementados, treinados e testados. O intuito é demonstrar com experiência prática que os modelos de AM e AP proporcionam resultados mais competitivos do que os modelos econométricos.

1.1 OBJETIVOS

Tendo como grande desafio a análise de abordagens para previsão de séries temporais financeiras, o escopo deste trabalho concentra-se nos seguintes objetivos.

1.1.1 Geral

Comparar as principais técnicas de predição para ST no contexto de mercado financeiro.

1.1.2 Específicos

1. Fazer uma análise qualitativa sobre o estado da arte sobre na predição de ST e teorias de previsão nos mercados financeiros;
2. Definir os processos de obtenção e preparação dos dados;
3. Definir os algoritmos mais adequados a serem implementados visando as áreas econométrica, AM e aprendizado profundo;
4. Criar modelagens computacionais para as técnicas escolhidas no item anterior;
5. Treinar os modelos escolhidos;
6. Fazer uma análise comparativa dos resultados dos preditores;
7. Desenvolver um repositório e disponibilizar na internet, de maneira a tornar todos os resultados desta pesquisa amplamente reproduzíveis.

1.2 ESCOPO DO TRABALHO

O escopo deste projeto se delimita no estudo comparativo de técnicas econométricas, AM e AP para a previsão de ST financeiras.

Será analisado o histórico de variação de preço da empresa de capital aberto com maior participação relativa no índice Bovespa, conforme tabela 01.

Ação	Participação Percentual Relativa
VALE	10,155
ITAUUNIBANCO	7,414
BRADESCO	5,612
PETROBRAS	5,610
B3	5,405

Tabela 01. As cinco empresas com maior participação no índice Bovespa³ durante o período de maio a agosto de 2020.

Não está no escopo do trabalho:

- Aplicar técnicas de aprendizado não supervisionado e por reforço
- Fazer análise fundamentalista sobre os papéis selecionados
- Criar um sistema automatizado de negociação
- Criar novos algoritmos para predição de ST
- Modelar e implementar softwares que fazem extração de dados para recuperar quaisquer informações de websites públicos.

1.3 MÉTODO DE PESQUISA

O presente estudo se caracteriza como uma pesquisa quati-quali (GIL, 2010). No que se refere a pesquisa qualitativa, busca-se compreender o que está sendo considerado o estado da arte do AM na predição de ST financeiras. Para isto, será feito um levantamento das publicações que envolvem os seguintes assuntos:

- a) Fundamentos de ST e AM;
- b) Predição de ST utilizando AM e modelos estatísticos.

³ Fonte: <http://bvmf.bmfbovespa.com.br/indices/ResumoCarteiraQuadrimestre.aspx?Indice=IBOV&idioma=pt-br>

Para abordagem quantitativa da pesquisa será feita a aplicação das técnicas de ST. Os dados serão obtidos em formato csv através da plataforma MetaTrader que trás informações de valores de mínimo, máximo, abertura e fechamento do preço de ações numa granularidade diária.

Para manipulação, pré-processamento, preparação, exploração estatística, divisão, treinamento, backtesting e avaliação dos resultados será utilizado bibliotecas que auxiliam a implementação das modelagens matemáticas. Na sequência, os resultados ficaram expostos em uma tabela trazendo as métricas selecionadas para comparação de cada algoritmo.

Os meios empregados neste projeto serão todos de código livre, ou seja, que possuem licença open source.

Todo o desenvolvimento utilizará a linguagem Python, (Python, 2021) e Jupyter Notebooks⁴ para exibição dos resultados.

A disponibilização dos código e dos resultados ficará em um servidor remoto de Git, o Github. Ademais, esta ferramenta servirá para controle de versão de código. O acesso ao material poderá ser visualizado através deste link: <https://github.com/brunocampos01/tcc-ufsc-grad>

1.4 ORGANIZAÇÃO DO TRABALHO

Este Trabalho de Conclusão de Curso está estruturado em 6 capítulos. No capítulo 1 é feito uma introdução, descrito o objetivo geral, específico, escopo do trabalho e também o método de pesquisa. No capítulo 2 são apresentados os principais conceitos referenciados neste trabalho. No capítulo 3 são mostrados os trabalhos que fundamentam esta pesquisa. A preparação, desenvolvimento, testes das modelagens e análise comparativa dos resultados são realizadas no capítulo 4. Por fim, no último capítulo, são enunciados as considerações finais deste trabalho de conclusão de curso.

⁴Endereço da biblioteca: <https://jupyter.org/>

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Séries Temporais

Segundo Sérgio (2017) uma ST pode ser definida como um conjunto de observações coletadas ao longo do tempo, em uma determinada ordem. Do ponto de vista estatístico, uma série de dados históricos pode ser tratada como uma sequência de variáveis aleatórias. Uma ST pode então ser referida como um processo estocástico discreto ao longo do tempo. Cada dado coletado pode ser um ponto em um gráfico de duas dimensões, onde o eixo das ordenadas determina as medições dos dados e o eixo das abscissas delimita em que momento discreto do tempo tais medições foram aferidas.

Na Figura 02 é apresentado um exemplo de uma ST com dados da variação de preço das ações da VALE3 em um período de 10 anos.

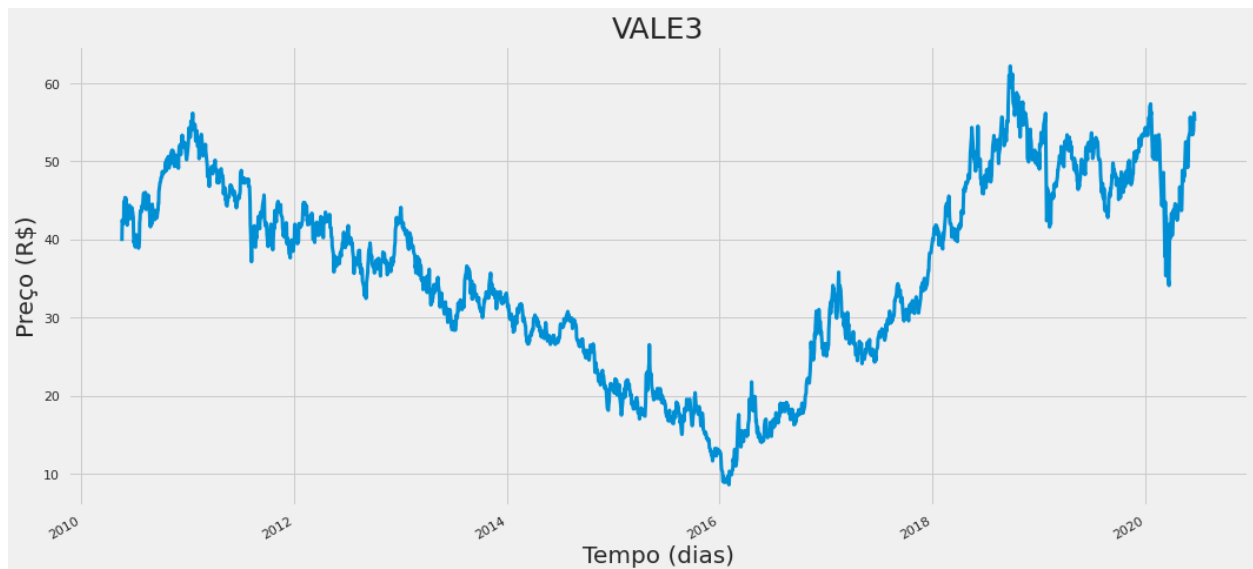


Figura 02. ST do papel VALE3⁵.

⁵ Figura desenvolvida pelo próprio autor.

2.1.1 Componentes de Séries Temporais

Para um melhor entendimento das ST é importante realizar a decomposição destas. As características mais importantes de uma ST são a tendência, sazonalidade e resíduo, (Cowpertwait; Metcalfe, 2009).

A Tendência (T) é um movimento regular ao longo prazo podendo assumir uma grande variedade de padrões, por exemplo, linear, exponencial, amortecido, e polinomial (Cowpertwait; Metcalfe, 2009).

A sazonalidade (S) é a ocorrência de padrões cíclicos de variação que se repetem, a um ritmo relativamente constante, juntamente com o componente de tendência (Cowpertwait; Metcalfe, 2009).

Os Resíduos (R) são as flutuações de curto prazo que não são sistemáticas nem previsíveis, (Cowpertwait; Metcalfe, 2009).

2.1.2 Decomposição

A decomposição pode ser aditiva ou multiplicativa, (Cowpertwait; Metcalfe, 2009). Na decomposição aditiva, os dados totais são tomados como a soma dos padrões decompostos. Conforme demonstrado na equação 1.

$$Z_t = \text{sazonalidade } (S_t) + \text{tendência } (T_t) + \text{ruído } (R_t) \quad 1$$

A decomposição aditiva é eficaz quando os valores de pico dos dados sazonais não variam muito. Na Figura 03 há um exemplo da ST e os seus elementos decompostos de forma aditiva.

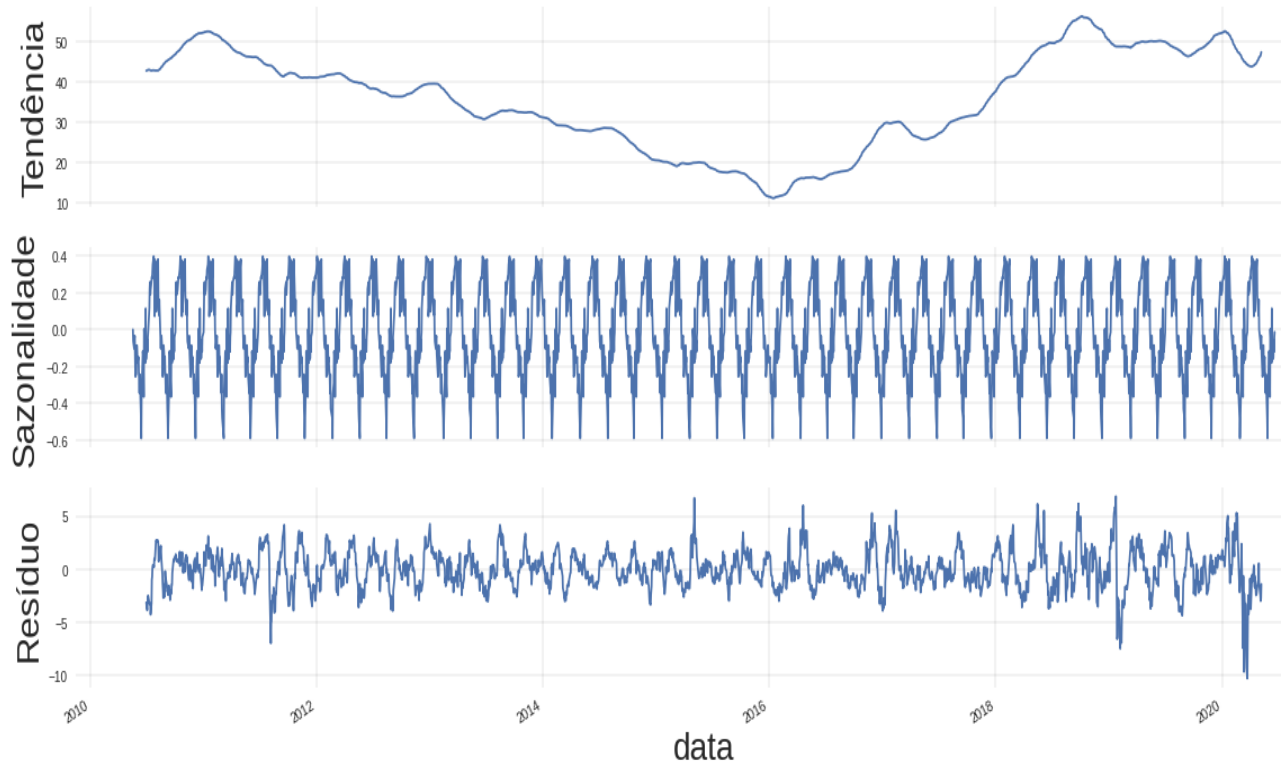


Figura 03. Decomposição aditiva da ST do papel VALE3⁶.

Já na decomposição multiplicativa, os dados da ST são tratados como o produto dos padrões decompostos, conforme demonstrado na equação 2.

$$Z_t = \text{sazonalidade } (S_t) \times \text{tendência } (T_t) \times \text{ruído } (R_t) \quad 2$$

A decomposição multiplicativa é eficaz quando o valor sazonal muda ao longo do tempo. Na Figura 03 há um exemplo da ST e os seus elementos decompostos de forma multiplicativa.

⁶ Figura desenvolvida pelo próprio autor.

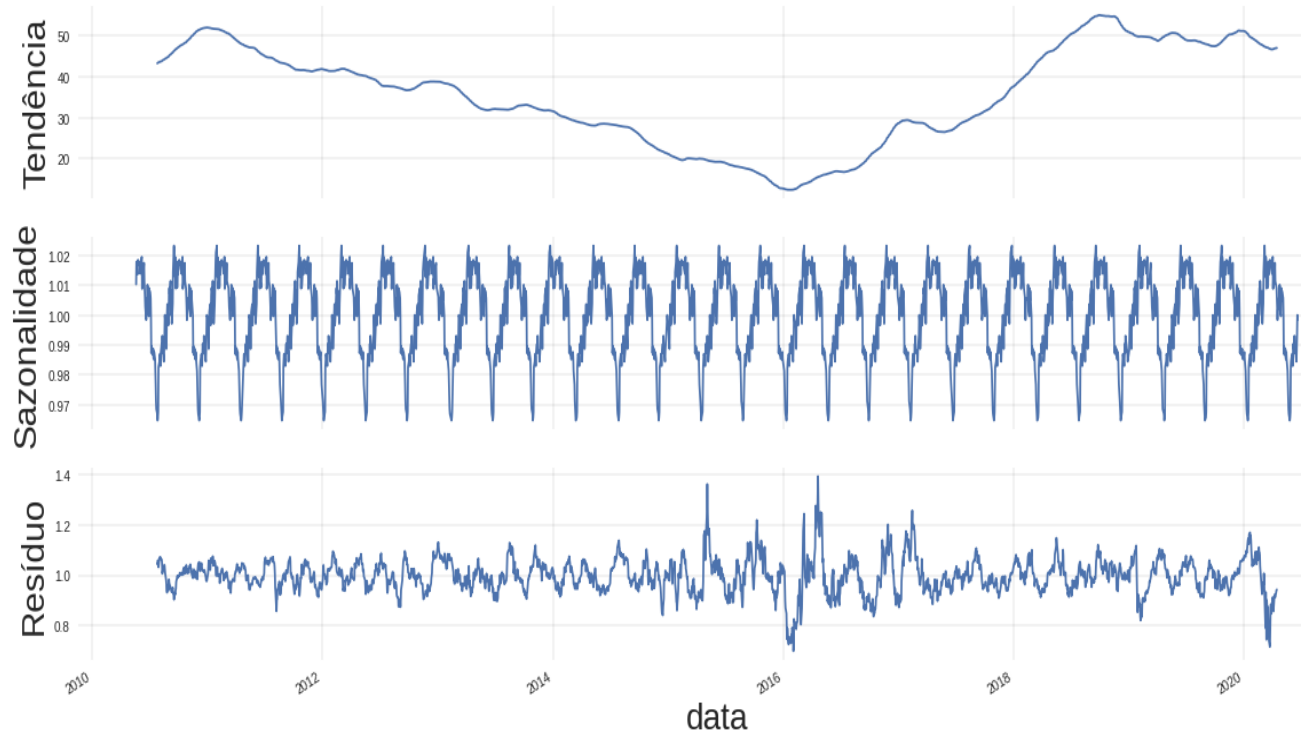


Figura 04. Decomposição multiplicativa da ST do papel VALE37.

2.1.3 Estacionariedade

Conforme mencionado em [Manav Kaushik et al, \(2020\)](#), uma característica central no desenvolvimento de modelos de ST com uma abordagem econométrica é uma suposição de alguma forma de equilíbrio estatístico, conhecido como estacionariedade.

O conceito de estacionariedade é a principal ideia que se deve ter para estimar uma ST. É fundamentalmente a constatação de estacionariedade que permitirá proceder inferências estatísticas sobre os parâmetros estimados com base na realização de um processo estocástico ([Bueno, 2011](#)).

2.1.4 Teste de Estacionariedade

Ao analisar os dados de uma ST é preciso garantir que estes são estacionários, ou seja, que a ST possui um comportamento ao longo do tempo e que possui uma alta probabilidade de seguir este mesmo comportamento no futuro. Em econometria, o teste de estacionariedade,

⁷ Figura desenvolvida pelo próprio autor

também conhecido como Augmented Dickey-Fuller (ADF) é uma técnica para testar a hipótese da estacionariedade da ST, [Manav Kaushik et al, \(2020\)](#).

2.2 PREVISORES ESTATÍSTICOS

O modelo de [BOX et al. \(2015\)](#), também conhecido como auto-regressivo integrado de média móvel, ARIMA (AutoRegressive Integrated Moving Average), foi elaborado a partir da coletânea de diversos trabalhos publicados na época, trazendo como mérito, a unificação das principais técnicas estatísticas, em uma única metodologia para construir modelos de predição de ST.

Conforme mencionado em [Parmezan \(2016\)](#), a seleção de modelos ARIMA é apropriada quando a série em estudo é estacionária, ou seja, suas propriedades estatísticas básicas, como média, variância e covariância, permanecem constantes ao longo do tempo.

De acordo com [BOX et al. \(2015\)](#), qualquer processo estocástico pode ser representado por um modelo auto-regressivo ou médias móveis. No caso de processos estocásticos não estacionários, aplica-se o operador de diferenciação para tornar a série estacionária.

Os modelos ARIMA envolvem três parâmetros estatísticos: autorregressão, integração e médias móveis. A parte auto regressiva (AR) indica que a variável de interesse pode ser obtida a partir de seus próprios atrasos. O termo integrado (I) indica a inclusão da diferença entre os valores atuais e os valores passados, sendo que tal diferenciação pode ser realizada mais de uma vez. O objetivo da integração é indicar o número de diferenças necessárias para garantir a estacionariedade da ST. Por fim, a parte de médias móveis (MA) utiliza os últimos valores históricos da ST para prever o próximo valor. O número de observações é sempre constante pois visa analisar a autocorrelação dos resíduos atuais com o que ocorreu no passado. A ordem de elementos dos modelos ARIMA são:

- $ARIMA(p, 0, 0) = AR(p)$;
- $ARIMA(0, 0, q) = MA(q)$;
- $ARIMA(p, 0, q) = ARMA(p, q)$

O modelo que explora a autocorrelação sazonal é conhecido como SARIMA. Ele contém uma parte não sazonal com parâmetros (p,d,q) , provinda do ARIMA, e uma sazonal com parâmetros (P,D,Q) , (Cowpertwait; Metcalfe, 2009). A aplicação do SARIMA é apropriada, sobretudo, em cenários nos quais os dados possuem variações sazonais.

2.3 PREVISORES DE APRENDIZADO DE MÁQUINA

De acordo com Parmezan (2016), os métodos de AM para previsão, em oposição aos modelos estatísticos, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos. Devido a isso, são mais simples de serem ajustados e alcançam considerável desempenho mesmo quando aplicados à ST complexas e altamente não-lineares. Modelos de AM são orientados a dados, o que significa dizer que, a princípio, não há necessidade de entender o conjunto de dados no qual o algoritmo será treinado. Por exemplo, dada uma ST, não é preciso decidir de antemão a respeito da linearidade ou não do modelo pois os algoritmos são responsáveis pelo ajuste automático. Essa característica é uma vantagem do AM frente a modelos estatísticos.

2.3.1 Máquina de Suporte de Vetores

O algoritmo de Máquina de Suporte de Vetores (SVM), é um modelo de AM baseado na Teoria da Aprendizagem Estatística, proposto por Vapnik (1999). Conforme mencionado em Sérgio (2017), o objetivo das SVM é minimizar o risco estrutural através de um subconjunto de padrões de dados de treinamento, chamados vetores de suporte. É possível citar como vantagens do uso de SVM sua alta capacidade de generalização, robustez para categorização de dados com dimensões altas e teoria bem estabelecida nas áreas de matemática e estatística.

Referente ao conceito de generalização das SVM pode ser exemplificado com uma classificação binária. A partir de duas classes e um conjunto de dados, as SVM determinam o

hiperplano que os separa, de maneira a colocar a maior quantidade possível de pontos da mesma classe do mesmo lado. Na Figura 05 é ilustrado este exemplo.

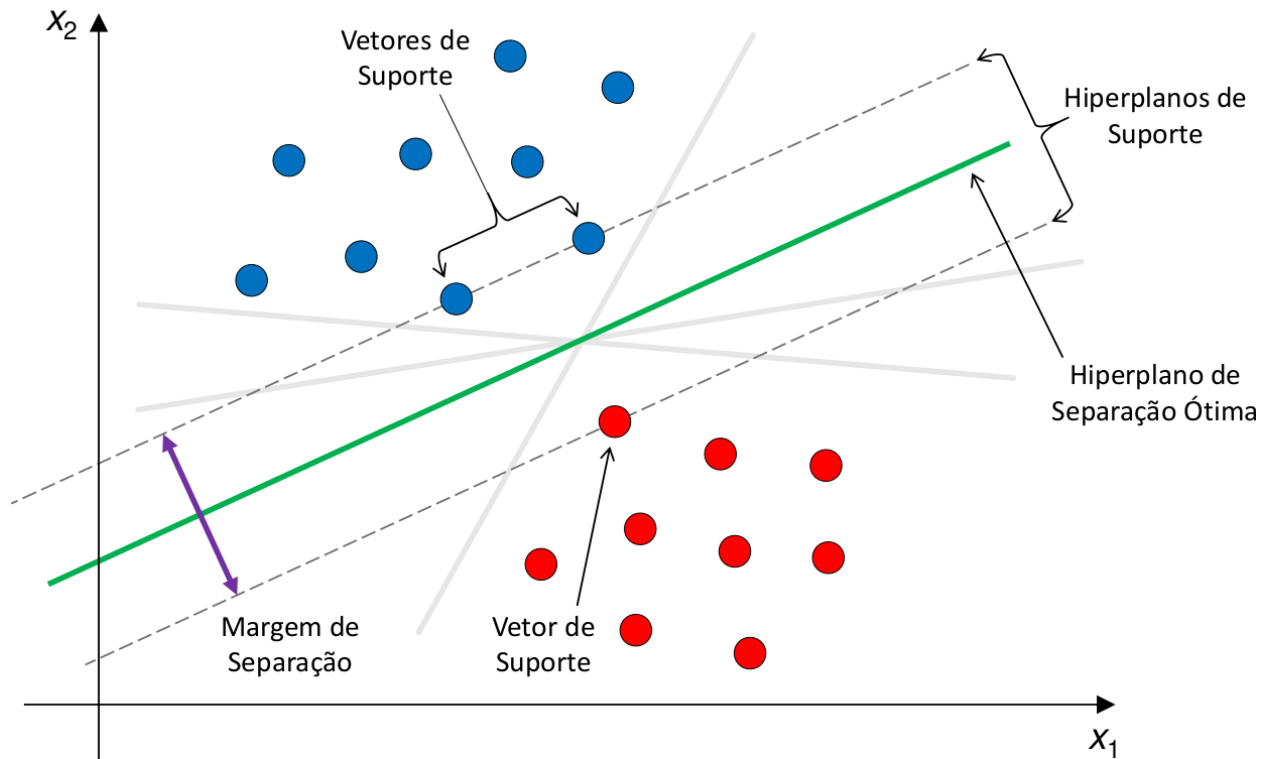


Figura 05. Hiperplano de separação ótima e seus hiperplanos de suporte. Os eixos ordenados x_1 e x_2 representam as dimensões das amostra no espaço 2D⁸.

Além de problemas de classificação, as SVM podem resolver problemas de regressão, a qual são chamadas de SVR.

No trabalho de [Manav Kaushik et al, \(2020\)](#), que tem por objetivo fazer a previsão de ST financeiras, é implementado foi uma SVM regressiva (SVR). O fato desta escolha se deve a uma abordagem de minimização de riscos estruturais que uma SVR proporciona, onde procura minimizar um limite superior do erro de generalização em vez de minimizar o erro de treinamento ao contrário da maioria das técnicas tradicionais de AM que adotam o risco empírico. Isto resulta em melhor generalização em relação às técnicas convencionais. Como

⁸ Fonte: [Parmezan \(2016\)](#)

ponto negativo, o autor destaca que as SVMs podem apresentar baixo desempenho se o conjunto de dados de treinamento for grande ou muito ruidoso, [Manav Kaushik et al, \(2020\)](#).

2.4 PREVISORES DE APRENDIZADO PROFUNDO

As Redes Neurais Artificiais (RNA) representam uma abordagem em inteligência artificial (IA) e suas constituições simulam a estrutura de funcionamento dos neurônios no cérebro humano [\(Luger, 2013\)](#). Uma RNA, similarmente ao seu paralelo biológico, é composta por determinado número de neurônios conectados por conexões sinápticas. Cada sinapse possui um peso atribuído, onde o conhecimento acumulado na rede depende, diretamente, da atualização desses pesos. Cada neurônio, ou também chamado de perceptron, possui um conjunto de entradas e de saídas ligadas a outros neurônios. Um neurônio possui um patamar de ativação que deriva da conjugação das entradas atuais recebidas com os pesos sinápticos das conexões, cujo valor resultante poderá ativar ou não a saída desse neurônio, [\(Luger, 2013\)](#).

2.4.1 Uso de RNA para a previsão de ST

Segundo [Manav Kaushik et al, \(2020\)](#), nos últimos anos, as RNA ajudaram a analisar com mais precisão as ST. Uma RNA é eficaz no mapeamento da dinâmica das ST não estacionárias dadas as propriedades adaptativas. As redes neurais são conhecidas por causa das suas funções de aproximação que podem mapear qualquer função não-linear sem quaisquer pressupostos prévios sobre os dados.

Como ponto de atenção para a previsão de ST, os modelos de RNA enfrentam o problema de sobreajustamento e erro de decaimento da retropropagação para previsão de ST, [\(Manav Kaushik et al, 2020\)](#).

Para resolver os problemas citados anteriormente é possível utilizar uma arquitetura recorrente. As RNA que utilizam uma arquitetura recorrente se chamam Redes Neurais Recorrentes (RNN). Para a previsão de ST a RNN mais utilizada é a Memória de Longo e Curto

Prazo (LSTM) onde são capazes de captar a não linearidade e a aleatoriedade de dados de ST de forma mais eficaz do que uma RNA mais simples, bem como superar o problema de decadência do erro de retropropagação através da memória de blocos de LSTM, e assim mostra capacidades superiores para predição de ST com dependência temporal longa, (Manav Kaushik et al, 2020).

Com a capacidade de memorizar dados históricos longos e determinar automaticamente o erro de decaimento da retropropagação, as RNA com arquitetura de LSTM atingem melhores resultados de previsão para problemas de ST em comparação aos modelos mais simples de RNA.

2.5 AVALIAÇÃO DOS RESULTADOS DE PREVISÃO

Conforme mencionado em Parmezan (2016), a avaliação do desempenho preditivo por meio do emprego das medidas descritas é tão essencial quanto a análise da distribuição gráfica dos erros de predição.

Entre as medidas extensivamente usadas para calcular os erros de previsão está o Erro Médio Percentual Absoluto (MAPE, Mean Absolute Percent Error), mostrado na fórmula 01.

$$MAPE = \frac{1}{h} \sum_{t=1}^h |z_t - \hat{z}_t| \times 100 \quad 1$$

O retorno da equação acima é um valor percentual que relaciona o valor previsto com o valor real da ST. Embora essa medida possibilita a comparação de erros entre sequências de dados com escalas diferentes, ela abrange uma deficiência prática: se uma ST possuir valores zero, ocorrerá uma imprópria divisão por zero, Parmezan (2016).

O MSE, não apresenta esse problema e, devido a isso, pode ser uma alternativa à medida MAPE. A fórmula do MSE é apresentada na fórmula 02.

$$MSE = \frac{1}{h} \sum_{t=1}^h |z_t - \hat{z}_t|^2 \quad 2$$

Na fórmula da MSE, a soma quadrática dos erros de predição é dividida pela quantidade de observações investigadas. O resultado esperado é que a soma de todos os erros computados seja próxima do valor zero.

Tanto o MSE quanto o MAPE são úteis, por exemplo, para comparar dois modelos preditivos.

3 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados, discutidos e analisados trabalhos que possuem objetivos semelhantes ao deste.

3.1 Revisão Sistemática

Para realização do Trabalho de Conclusão de Curso foi utilizado como recurso metodológico a Revisão Sistemática, pois, conforme [Sampaio e Mancini, \(2007\)](#), "uma revisão sistemática, assim como outros tipos de estudo de revisão, é uma forma de pesquisa que utiliza como fonte de dados a literatura sobre determinado tema. Esse tipo de investigação disponibiliza um resumo das evidências relacionadas a uma estratégia de intervenção específica, mediante a aplicação de métodos explícitos e sistematizados de busca, apreciação crítica e síntese da informação selecionada", ([Sampaio e Mancini, 2007](#)).

Tendo como objetivo prévio a definição do tema "Previsão de ST Financeiras" a estratégia específica de investigação foi a utilização dos seguintes termos norteadores: machine learning stock market; machine learning time series; financial time series; financial data processing. Estes termos de pesquisa foram elaboradas com base no know-how prévio do autor e através do aprimoramento durante o estudo das publicações investigadas no decorrer da exploração, a qual foi executada a partir de julho de 2019 e com o suporte dos motores de busca Arxiv, Dl.acm.org, Hindawi, Ieeexplore, Onlinelibrary.wiley, Researchgate, Sciencedirect e Springer.

Os resultados da pesquisa bibliográfica podem ser observados na tabela 02.

Pesquisa Bibliográfica				
Site	machine learning stock market	machine learning time series	financial time series	financial data processing
ieeexplore	399	2.607	592	3.363
sciencedirect	1	72	3.570	39
researchgate	2	4	349	23
arxiv	156	161	405	203
hindawi	0	4	288	3
onlinelibrary.wiley	66	3	125	3
springer	8	283	11	383
dl.acm.org	0	8	145	10

Tabela 02. Resultados da pesquisa bibliográfica.

Após execução da pesquisa citada acima foi observado as seguintes características:

- Aprendizado por transferência não foi encontrado resultados;
- Aprendizado por reforço foi encontrado poucos resultados;
- A maioria dos resultados utilizam a técnica de LSTM e SVM

E por fim, foi definido como estratégia a exploração de um histórico de técnicas já utilizadas na área econométrica e comparada com técnicas de aprendizado de máquina e aprendizado profundo. Os artigos definidos para a base teórica do trabalho eram os que fizeram comparativos entre as técnicas, filtrando da pesquisa bibliográfica 38 trabalhos científicos e utilizando a partir da proposta definida o recurso de 12 referenciais científicos para o desenvolvimento e conclusão deste trabalho.

3.2 Predição de Séries Temporais por Similaridade

[Parmezan \(2016\)](#), em seu trabalho, tem como principal objetivo a comparação de preditores de ST por similaridade com métodos estatísticos. Diante disso, é gerada a seguinte hipótese:

“Métodos para predição de Séries Temporais baseados em similaridade podem prover resultados competitivos em relação aos obtidos com a aplicação de métodos estatísticos estado-da-arte.”

A hipótese levantada pelo autor se baseia em premissas como a fácil compreensão de métodos por similaridade e também, que algoritmos baseados em grau de semelhança tem eficácia na extração de padrões temporais.

Os algoritmos que implementam métodos estatísticos assumem que os dados seguem alguma distribuição conhecida e, a partir disso, definem parâmetros de funções para ajustar um modelo aos dados. Contudo, a utilização desse tipo de abordagem, também chamada de paramétrica, acaba se tornando um limitador por envolver alta complexidade matemática e exigir vasto conhecimento técnico para o estabelecimento dos parâmetros do modelo. Em termos práticos, definir os valores dos parâmetros de um modelo estatístico reside em quatro etapas ([BOX et al., 2015](#)):

1. Seleção da estrutura do modelo conforme as características dos dados;
2. Identificação das ordens do modelo;
3. Estimação dos coeficientes do modelo;
4. Diagnóstico do modelo ajustado.

Como resultado da dificuldade em se estimar modelos estatísticos, diversos estudos vêm sendo empreendidos no intuito de criar uma modelagem não-paramétrica para a predição de ST. Uma das principais vantagens da abordagem não-paramétrica é que esta não pressupõe sobre a natureza da distribuição dos dados.

Para construção de modelos preditivos de ST, o autor divide em grupos conforme exibido na Figura 06.

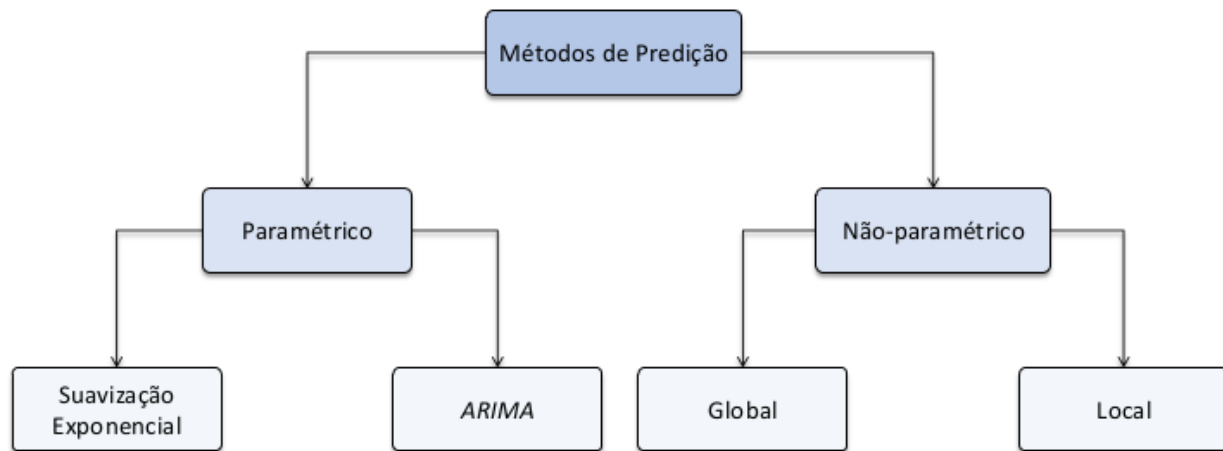


Figura 06. Hierarquia de abordagens para predição de ST⁹.

- **Métodos Paramétricos:** Os métodos estatísticos se distinguem por utilizar, na confecção do modelo de predição, o conhecimento a priori sobre a natureza da distribuição dos dados. Esse controle probabilístico faz com que o modelo dependa explicitamente de um conjunto finito de parâmetros, os quais devem ser determinados de maneira a otimizar os resultados da predição.
- **Métodos Não-Paramétricos:** Os métodos de AM para predição, em oposição aos modelos estatísticos, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos. Por não dependerem explicitamente de parâmetros para modelar o comportamento do fenômeno, esses métodos são mais simples de serem ajustados e demonstram considerável desempenho mesmo quando aplicados a séries complexas e altamente não-lineares.

[Parmezan \(2016\)](#), em seu trabalho, contempla um capítulo para explicar a similaridade entre ST. Segundo o autor, a similaridade pode ser entendida como uma estimativa do grau de semelhança entre dois objetos. Na predição de ST utilizando AM, quantificar quão parecidas são

⁹ Fonte: [Parmezan \(2016\)](#)

duas subsequências de dados para decidir se elas pertencem a um mesmo espaço de características é uma tarefa altamente subjetiva e que sofre influência de distintos fatores.

O autor conclui que o interesse da dissertação incidiu no fato de encontrar uma importante lacuna de pesquisa. Essa lacuna refere-se à comparação objetiva e subjetiva entre modelos estatísticos e de AM para a predição de ST. A partir disso, o trabalho abrangeu duas temáticas principais: (1) a exploração das propriedades inerentes à predição baseada em similaridade e (2) a comparação do algoritmo k-Nearest Neighbors - Time Series Prediction with Invariances (kNN-TSPI), aqui proposto, com os métodos estatísticos e de Aprendizado de Máquina considerados o estado-da-arte na literatura.

3.2 Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques

A previsão de dados de ST, obedece uma hipótese geral onde o comportamento passado da ST contém todas as informações necessárias para prever seu comportamento futuro. Assim, tradicionalmente, a maioria das tentativas para prever ST financeiras entregou modelos estatísticos. No entanto, o maior problema com modelos estatísticos como ARIMA¹⁰, pode ser resumido em duas grandes suposições:

1. A ST prevista é linear
2. A ST que está sendo prevista é estacionária

Ademais, não levam em conta os efeitos de outros parâmetros que podem ser cruciais ao determinar o valor futuro de uma variável macroeconômica.

No trabalho de [Manav Kaushik et al, \(2020\)](#), é implementado um modelo estatístico chamado VAR, que é comumente utilizado como referência para a análise e previsão multivariada de ST. Como ponto negativo deste modelo, o autor destaca que, o VAR (Vector autoregression) geralmente falha no mapeamento de uma associação não-linear entre diferentes

¹⁰ Autoregressive Integrated Moving Average

variáveis e, como consequência, tem baixa generalização. Além disso, como o modelo ARIMA, o modelo VAR também requer uma série de dados estacionários.

O modelo de AM implementado é uma SVM. O fato desta escolha se deve a uma abordagem de minimização de riscos estruturais que uma SVM proporciona, onde procura minimizar um limite superior do erro de generalização em vez de minimizar o erro de treinamento ao contrário da maioria das

técnicas tradicionais de AM que adotam o risco empírico. Isto resulta em melhor generalização em relação às técnicas convencionais. Como ponto negativo, o autor destaca que as SVMs podem apresentar baixo desempenho se o conjunto de dados de treinamento for grande ou muito ruidoso.

O terceiro modelo implementado por [Manav Kaushik et al, \(2020\)](#), é uma RNA com arquitetura recorrente. As RNAs são eficazes no mapeamento da dinâmica de ST não estacionárias e com ruídos. Os modelos de uma única camada, os perceptrons, podem enfrentar o problema de erro de sobreajustamento, por isso, o autor apresenta uma abordagem de RNA com arquitetura recorrente, usando memórias de longo e curto prazo. Esse tipo de RNA recorrente se chama LSTM e tem como diferencial poder de captar a aleatoriedade e não-linearidade dos dados de ST de forma mais eficaz. Com a capacidade de memorizar longos dados históricos e determinam automaticamente o tempo ótimo de atraso, as LSTMs atingem maior precisão de previsão e se generalizam bem com diferentes intervalos de previsão.

A preparação dos dados começa selecionando as variáveis macroeconômicas que mais influenciam a variação da ST financeira. O artigo em questão tem como objetivo prever a taxa de câmbio entre a moeda americana com a moeda indiana, por isso são utilizados as seguintes informações:

- Índice de Preços no Consumidor (IPC)
- Índice de Produção Industrial (IIP)
- Taxas de Juros
- Oferta Monetária
- Reservas Totais
- Índice da bolsa de valores

- Comércio (Exportações Líquidas)

Em seguida, é calculada a diferença entre os valores de cada variável macroeconômica dos respectivos países. Para analisar a causalidade entre ST é aplicado o teste de hipótese de Granger. Com o resultado do teste, é possível evidenciar que todas as variáveis resultantes da etapa anterior apresentam causalidade na taxa de câmbio. Para verificar a estacionariedade das ST, é aplicado o teste de hipótese Dickey-Fuller aumentado.

Para a escolha dos modelos que apresentam melhor performance, é utilizada as seguintes métricas:

- MAPE
- MPE
- RMSE
- Precisão

Na tabela 03 há os resultados da comparação entre os algoritmos selecionados.

Performance/Modelo	VAR	SVM	LSTM (RNA)
MAPE	0,0369	0,0283	0,0217
MPE	0,0286	-0,0261	0,0031
RMSE	2,9381	2,5585	1,6872
Precisão (%)	96,31%	97,17%	97,83%

Tabela 03. Comparação dos métodos de predição¹¹

Os autores concluem que a análise multivariada das séries temporais mostrou resultados satisfatórios. Modelos como ARIMA e VAR estabeleceram referências devido à sua ampla utilização na literatura, mas o AM e o AP superam os tradicionais modelos econômicos. O modelo de AP, LSTM mostrou o melhor desempenho para a previsão da taxa de câmbio USD/INR seguida de modelo AM, SVM, e depois pelo tradicional modelo econométrico, VAR.

¹¹ Fonte: [Manav Kaushik et al, \(2020\)](#)

3.3 Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques

No trabalho de [Patel et al, \(2015\)](#) são comparados três modelos de AM (SVM, floresta aleatória e Naive Bayes) e um de AP (RNA). O foco do trabalho em questão é o pré-processamento de dados com o objetivo de melhorar a precisão da previsão.

Para sustentar o argumento de previsão de ST financeiras, os autores citam a Hipótese de mercado eficiente de Fama (1970), a qual explica que os preços das ações são eficientes, o que significa que é possível prever os preços das ações com base nos dados de negociação. Portanto, se as informações obtidas com os preços das ações forem pré-processadas com eficiência e forem aplicados algoritmos apropriados, poderá ser prevista a tendência das ações ou do índice de preços das ações.

Desde anos, muitas técnicas foram desenvolvidas para prever tendências de ST financeiras. Inicialmente, métodos de regressão como ARIMA foram usados para prever tendências de ações. Como os dados de preço de ações são classificados sendo não estacionários foram utilizadas técnicas não lineares de AM.

Os parâmetros ideais são obtidos através de experimentação sobre 20% do total dos dados. Os 80% dos dados restantes, passam por uma divisão entre dados de treino e teste, sendo 50% para cada conjunto. Para iniciar o processo de preparação dos dados, são analisados indicadores técnicos financeiros, que são:

- Médias Móveis
- Stochastic K% , Stochastic D% e Larry Williams R%
- MACD
- RSI
- CCI
- Oscilador A/D
- Momentum

A partir destes, é feita uma normalização dos valores em uma camada com o nome de “Camada de preparação de dados determinísticos de tendência” para que estejam dentro do intervalo [-1,+1]. Depois da normalização são selecionados somente os valores máximo e mínimo de cada indicador técnico financeiro.

Para avaliar os resultados dos experimentos é proposto um comparativo utilizando as métricas de Precisão e medida F. Na tabela 04 há os resultados da comparação entre os algoritmos selecionados.

Ação/ Modelos				
	RNA	SVM	Floresta Aleatória	Bayes
S&P BSE SENSEX	Precisão: 0,7839 Medida F: 0,7849	Precisão: 0,7979 Medida F: 0,8168	Precisão: 0,8775 Medida F: 0,8794	Precisão: 0,7354 Medida F: 0,7547
NIFTY 50	Precisão: 0,8481 Medida F: 0,8635	Precisão: 0,8242 Medida F: 0,8438	Precisão: 0,9131 Medida F: 0,9178	Precisão: 0,8097 Medida F: 0,8193
Reliance Industries	Precisão: 0,6527 Medida F: 0,6786	Precisão: 0,7275 Medida F: ,7392	Precisão: 0,7420 Medida F: 0,7447	Precisão: 0,6565 Medida F: 0,6658
Infosys Ltda	Precisão: 0,7130 Medida F: 0,7364	Precisão: 0,7988 Medida F: 0,8119	Precisão: 0,8110 Medida F: 0,8176	Precisão: 0,7307 Medida F: 0,7446
Média	Precisão: 0,7494 Medida F: 0,7659	Precisão: 0,7871 Medida F: 0,8029	Precisão: 0,8359 Medida F: 0,8399	Precisão: 0,7331 Medida F: 0,7461

Tabela 04. Resultados dos métodos de predição por cada ação.¹²

¹² Fonte: Patel et al, (2015)

A precisão e a medida F relatadas são calculadas sobre os modelos com melhor desempenho. Pode-se observar que o processo Naive Bayes é o menos preciso, enquanto a Floresta Aleatória é a mais precisa, com precisão média de quase 84%.

A Figura 07 ilustra o processo de previsão de ST proposta pelos autores.

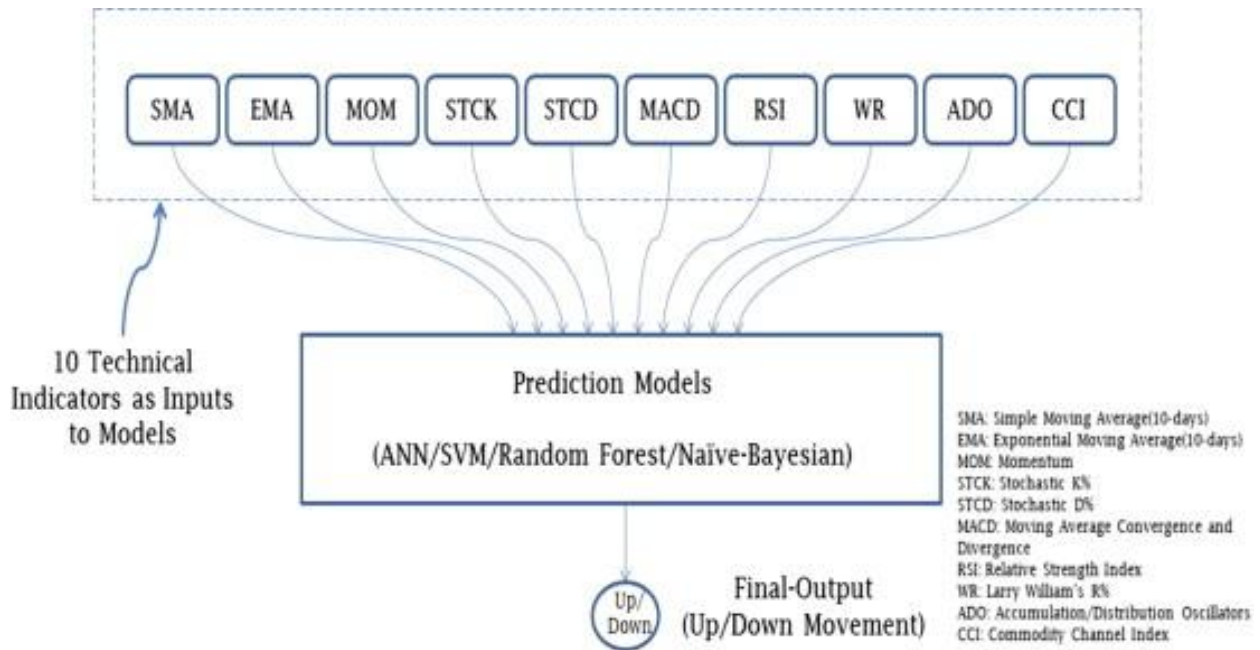


Figura 07. Processo de previsão proposto pelo autor¹³.

Além do capítulo de resultados experimentais, há um outro discutindo sobre os resultados obtidos. Um ponto a observar com base nos resultados é que os dados dos indicadores são preparados discretizando os valores contínuos. A ideia é baseada no fato de que cada parâmetro de valor, quando comparado com o valor do dia anterior, indica a tendência futura de aumento ou redução. Os dados são discretizados com base nessas heurísticas. Quando esses dados são fornecidos como entrada para o modelo, já está sendo inserido a tendência com base em cada parâmetro de entrada.

Os autores concluem que o sistema proposto pode ser implantado em tempo real para a previsão de tendências das ações, tornando os investimentos mais rentáveis e seguros.

¹³ Fonte: Patel et al, (2015)

3.4 Trabalhos Complementares com ênfase em Métodos Econométricos

O trabalho que é referência na área de previsão de ST financeiras usando métodos econométricos é o de (BOX et al., 2015). O trabalho em questão traz uma metodologia para definir os valores dos parâmetros de um modelo estatístico. Essa metodologia consiste em quatro etapas:

- Seleção da estrutura do modelo conforme as características dos dados;
- Identificação das ordens do modelo;
- Estimação dos coeficientes do modelo;
- Diagnóstico do modelo ajustado.

Todas essas etapas são guiadas por fundamentos da estatística descritiva e, na maioria dos casos, são realizadas de maneira semiautomática por meio do uso de funções baseadas em autocorrelação, cujos resultados podem ser interpretados via correlogramas, e da aplicação de técnicas para obtenção de argumentos de entrada a partir da minimização de critérios de informação, os quais penalizam o modelo pela quantidade de parâmetros suficientemente necessários para o seu ajustamento.

Um trabalho que traz a aplicação do método de Box-Jenkins é o artigo Junior, Salomon and Pamplona (2014). No artigo em questão é avaliado o desempenho do modelo ARIMA para ST no índice Bovespa. O método de investigação utilizado foi a modelagem matemática e seguiu o método Box-Jenkins. A comparação dos resultados com outros modelos foi realizada através do método MAPE (Average Absolute Percentage Error - Erro Percentual Médio Absoluto). Os resultados mostraram que o modelo utilizado obteve valores MAPE mais baixos, indicando assim uma maior aptidão. Como conclusão, é demonstrado que o modelo ARIMA pode ser utilizado para a previsão de ST relacionados com índice de mercado.

3.5 Trabalhos Complementares com ênfase em AM e AP

O trabalho de [Kaastra and Boyd, \(1996\)](#) tem servido como referência até os dias atuais por trazer um guia prático introdutório ao projeto de uma rede neural para a previsão de dados de ST econômicas. Constituído de 8 etapas que precisam ser observadas em modelos de previsão, além de trazer sugestões para a seleção das variáveis, coleta de dados, pré-processamento dos dados, treinamento e validação, topologia, critérios de avaliação de desempenho, parâmetros do algoritmo backpropagation e implementação.

Já [Daniel \(2019\)](#), traz técnicas para o processamento de dados de ST indo de encontro com o objetivo 2 deste Trabalho de Conclusão de Curso. Ademais, introduz os métodos de dimensionamento mais frequentes, depois compara a estacionaridade resultante e preservação de informação útil para a previsão de tendências. Também cita o método de divisão de dados específico das ST, evitando sobreajustamentos indesejados e propõe várias rotulagens para classificação e regressão. Como resultado o autor notou que a natureza financeira dos dados traz um processo estocástico o qual acrescenta um nível de complexidade no processamento. Portanto, é através da comparação de diferentes combinações de características e rótulos em cima das ST financeiras que mais impactam nos resultados, inclusive mais do que a afinação dos hiperparâmetros.

3.6 Considerações Finais

De acordo com [Parmezan \(2016\)](#), nas últimas duas décadas, com a utilização de processos para mineração de dados houve um aumento significativo no interesse em AM, especialmente regressão, para dar suporte à análise de fenômenos com dependência temporal. Mesmo assim, por causa da simplicidade e fácil interpretabilidade, os modelos estatísticos alcançaram o estado da arte para a modelagem e predição de ST.

Para [Manav Kaushik et al, \(2020\)](#), modelos como ARIMA e VAR (Vector autoregression) estabeleceram indicadores de referência econômica e financeira devido à sua ampla utilização, mas o AM e o AP superam claramente as tradicionais modelagens econométricas. Na análise realizada por [Manav Kaushik et al, \(2020\)](#), o modelo de AP, apresentou o melhor desempenho para a previsão da taxa de câmbio moeda entre a moeda americana com a moeda indiana, seguida de um previsor SVM, e depois pelo tradicional modelo econométrico, ARIMA.

Em [Patel et al, \(2015\)](#), não há uma comparação entre modelagens estatísticas com modelagens de AM. O foco do trabalho em questão é prever a direção do movimento. Contudo, o processo intermediário de modelagem faz comparações entre algoritmos de AM e AP que são RNA (Redes Neurais Artificiais), SVM, floresta aleatória e ingênuo-Bayes utilizando dados da bolsa indiana.

A tabela 05 analisa os trabalhos até então mencionados nesta sessão, de acordo com os métodos, métricas de avaliação e conjunto de dados.

	Métodos	Métricas	Fontes de dados
Parmezan (2016)	ARIMA, SARIMA, Suavização Exponencial, SVM, RNA	MAPE	S&P
Manav Kaushik et al, (2020)	ARIMA, VAR, SVM, RNA	MAPE, MPE, RMSE, acurácia	Câmbio USD/INR
Patel et al, (2015)	Bayes, SVM, Floresta Aleatória, RNA	Precisão, Medida F	S&P, NIFTY 50, Reliance Industries, Infosys Ltda

Tabela 05. Análise dos métodos, métricas e fontes de dados¹⁴

As pesquisas em predição de ST utilizam desde métodos estatísticos complexos, com uma quantidade considerável de parâmetros, até modelos intuitivos, simples e de fácil

¹⁴ Fonte: elaborada pelo autor.

implementação. Dessa forma, é necessário investigar e ponderar vantagens e desvantagens do uso de cada algoritmo nos campos estatísticos, de AM e AP.

Como contribuição, este trabalho se diferencia por fazer esta investigação do estado da arte sobre o tema e fazer uma comparação atualizada das técnicas para predição de ST financeiras.

4 DESENVOLVIMENTO

Tendo em vista os objetivos deste trabalho, será neste capítulo que ocorrerá a modelagem dos previsores estatísticos, de AM e AP.

Para os modelos econométricos será utilizado a metodologia de [BOX et al, \(2015\)](#). Este método constitui de quatro etapas detalhadas na figura 08.

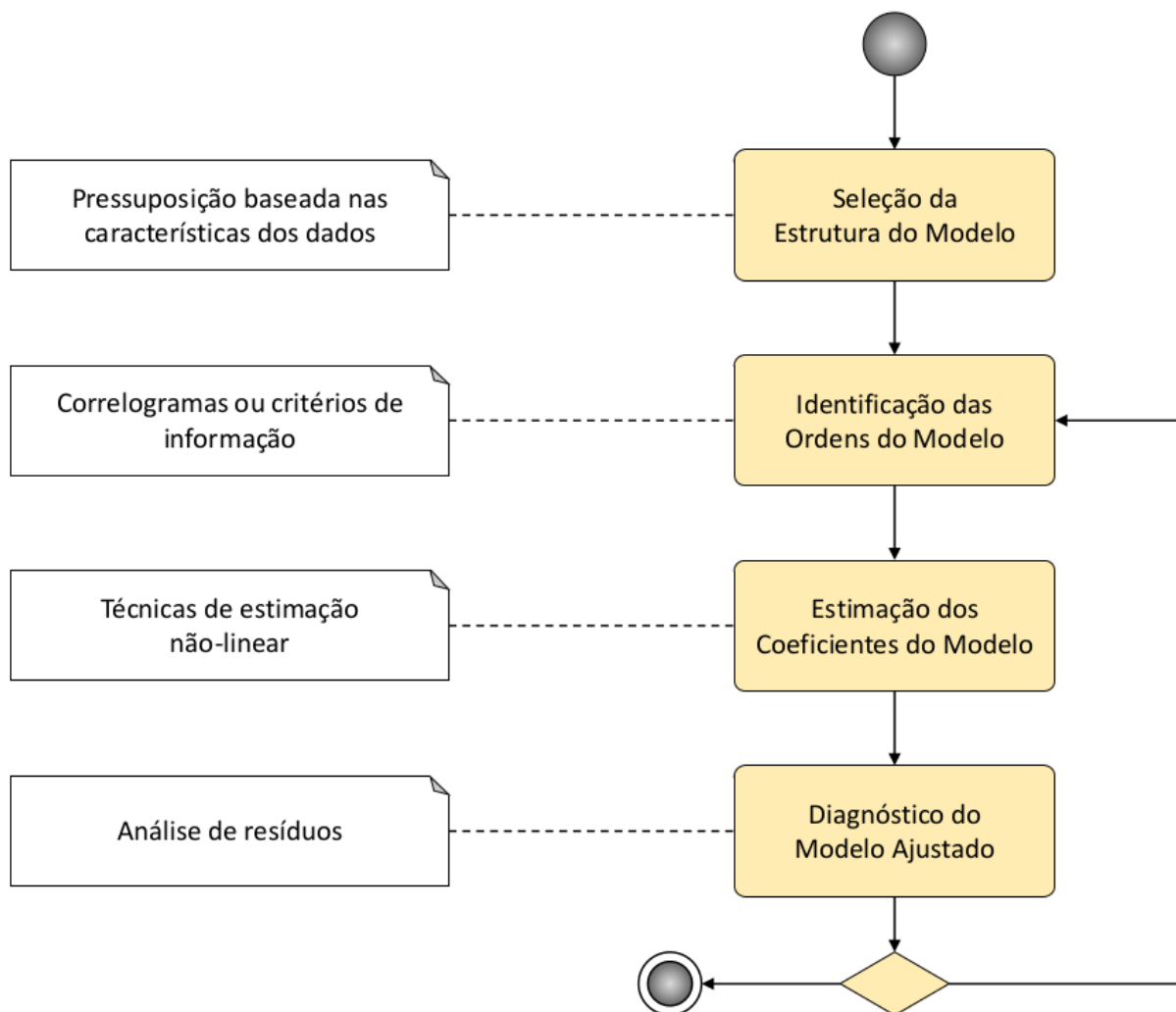


Figura 08. Ciclo iterativo de Box-Jenkins¹⁵.

¹⁵ Adaptado de [BOX et al, \(2015\)](#)

Já para as previsões de ST utilizando AM e AP será empregado a metodologia demonstrada em Parmezan (2016). A figura 09 ilustra o método:

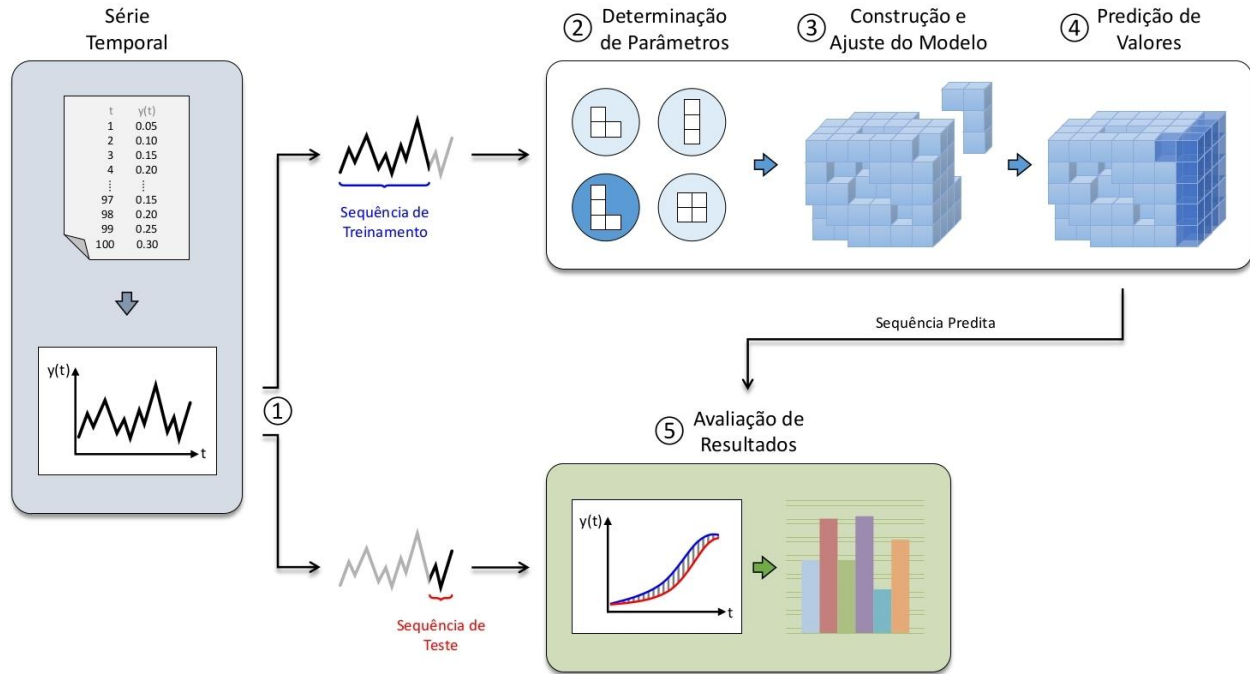


Figura 09. Etapas para realizar a previsão de ST¹⁶.

Primeira etapa: a ST é dividida em dados de treinamento e teste sendo que neste caso a primeira parte da ST serão dados de treinamento e a parte final dados de testes.

Segunda etapa: é feita a estruturação do modelo de previsão. Neste ponto é que são encontrados os melhores parâmetros empregando alguma técnica de busca. Usualmente, essa técnica é implementada por um algoritmo que recebe como entrada a sequência

Terceira etapa: a partir dos parâmetros identificados na etapa anterior, o modelo de interesse é construído e ajustado aos dados da sequência de treinamento.

Quarta etapa: é o momento que é feita a previsão de valores, ou seja, nesta etapa que são executados os modelos estatísticos, de aprendizado de máquina ou os modelos de aprendizado profundo.

¹⁶ Adaptado de Parmezan (2016)

Quinta etapa: é a etapa de avaliação do modelo. Nesta etapa é realizada a avaliação dos erros de previsão, ou seja, mensurar o quanto os valores da sequência predita se distanciam dos valores da sequência de teste.

4.1 Obtenção, Manipulação e Pré-processamento dos Dados

O MetaTrader¹⁷ é uma plataforma open source para negociação de ativos financeiros. A obtenção dos dados foi feita através da plataforma, a qual permitiu baixar uma tabela de dados no formato csv. Na tabela 06 há uma amostra dos dados obtidos.

date	price	open	high	low	volume	change %
2020-06-19	55.17	56.33	56.6	54.71	37.86	-1.78%
2020-06-18	56.17	56.3	56.54	55.61	25.69	-0.05%
2020-06-17	56.2	55.49	56.44	55.15	27.57	1.46%
2020-06-16	55.39	55.49	55.5	54.22	29.17	2.80%
2020-06-15	53.88	52.5	54.43	52.13	27.32	0.90%

Tabela 06. Amostra dos dados da VALE3, obtidos através do Metatrader.

A tabela de dados possui 2495 linhas e 7 colunas, sendo:

- Data (date), data que os dados foram gerados.
- Preço de fechamento (price), é o preço final, em reais, que a ação fechou o dia.
- Preço de abertura (open), é o preço de abertura, em reais, que a ação abriu o dia.
- Preço máximo alcançado no dia (high), é o preço máximo, em reais, que a ação chegou a alcançar no dia.

¹⁷ Endereço eletrônico da plataforma: <https://www.metatrader5.com/pt>

- Preço mínimo alcançado no dia (low), é o preço mínimo, em reais, q a ação alcançou o dia.
- Volume de ações negociadas (volume), é a quantidade de ações negociadas no dia.
- Mudança (change %), é porcentagem de alteração do preço em relação ao dia anterior

O período selecionado para a modelagem foi do dia 20 de maio de 2010 até 19 de junho de 2020, sendo uma granularidade diária de informação, ou seja, para cada dia foi obtido uma linha contendo as informações data, preço de fechamento, preço de abertura, preço máximo alcançado no dia, preço mínimo alcançado no dia, volume de ações negociadas e mudanças em percentual do preço. O total de linhas obtidas foi de 2495.

Para poder manipular as informações, foi utilizada uma estrutura de dados que se chama dataframe. O dataframe é uma estrutura de dados bidimensional no qual seus eixos são rotulados de linhas e colunas. As bibliotecas de Pandas¹⁸ e Apache Spark¹⁹ utilizam esta estrutura de dados para manipulação de dados e permitem trabalhar com os dados em memória RAM.

Em seguida foi feito um pré-processamento dos dados. A variável alvo, ou seja, a que se deseja prever é o preço de fechamento, por isso só foram mantidas as colunas de data e preço de fechamento.

4.2 Análise Exploratória da ST

O objetivo da exploração da ST é identificar padrões não aleatórios em cima de uma variável alvo e a partir disso poder fazer previsões sobre o futuro, orientando a tomada de decisões.

¹⁸ Endereço da biblioteca: <https://pandas.pydata.org/pandas-docs/stable/index.html>

¹⁹ Endereço da biblioteca: <https://spark.apache.org/>

4.2.1 Resumo dos Dados

Para começar foram analisadas as medidas de posição. Conforme mencionado em [Morettin e Bussab \(2010\)](#), as medidas de posição servem para resumir os dados através de um ou mais valores representativos para a ST. A tabela 07 traz as medidas calculadas.

Medidas	Preço R\$
média	35,80
mediana	36,88
moda	27,43
mínimo	8,60
máximo	62,20

Tabela 07. Medidas de posição dos dados do preço das ações da VALE3.

4.2.2 Características da Série Temporal

Uma ST possui componentes que não são possíveis de detectar pontualmente, por isso se faz necessário isolar essas características para uma análise melhor dos dados. Esta análise visa descrever os comportamentos da ST, tais como a existência ou não de tendência, variação sazonal e observações discrepantes (outliers). Na Figura 10, há a distribuição dos dados em relação ao tempo.

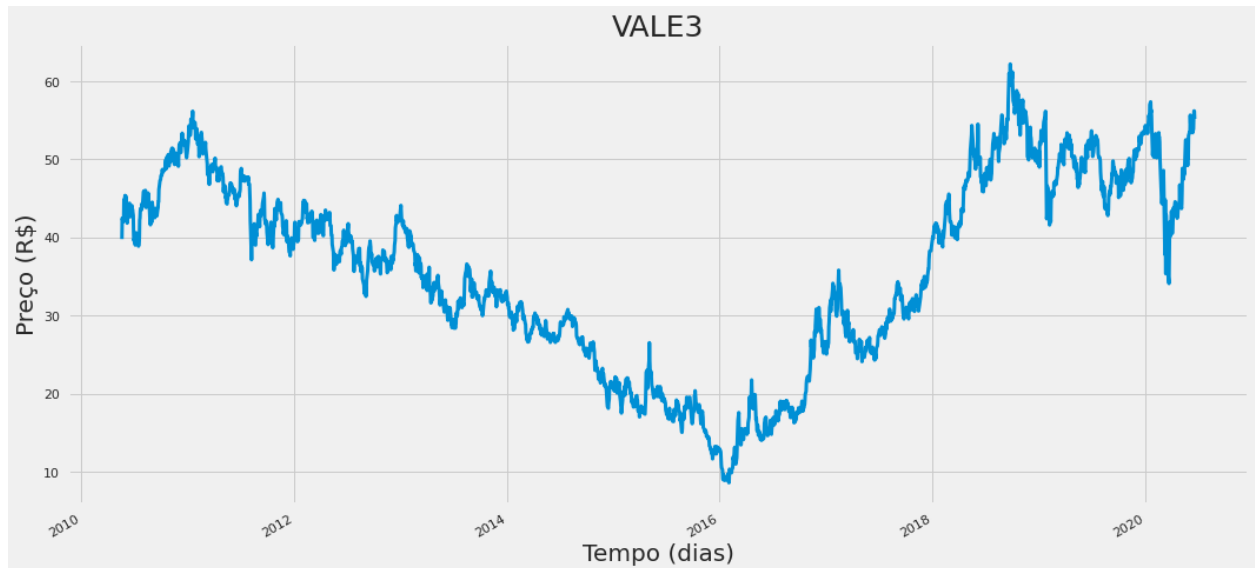


Figura 10. Gráfico de distribuição do preço de fechamento das ações da VALE3 em um período de 10 anos²⁰.

A partir deste gráfico é possível analisar que:

- Houve uma grande tendência de baixa até aproximadamente o ano de 2016.
- Depois de alcançar o fundo da tendência de baixa, houve reversão da tendência começando um ciclo de alta que durou aproximadamente 1000 dias.
- Nos últimos anos, a ST não apresenta uma tendência evidente.

Além da distribuição também é necessário observar a tendência, sazonalidade e resíduos da ST a fim de compreender o comportamento. Para analisar cada um destes componentes foi realizada a decomposição da ST. A Figura 11, traz um gráfico para cada uma destas características.

²⁰ Figura gerada pelo próprio autor.

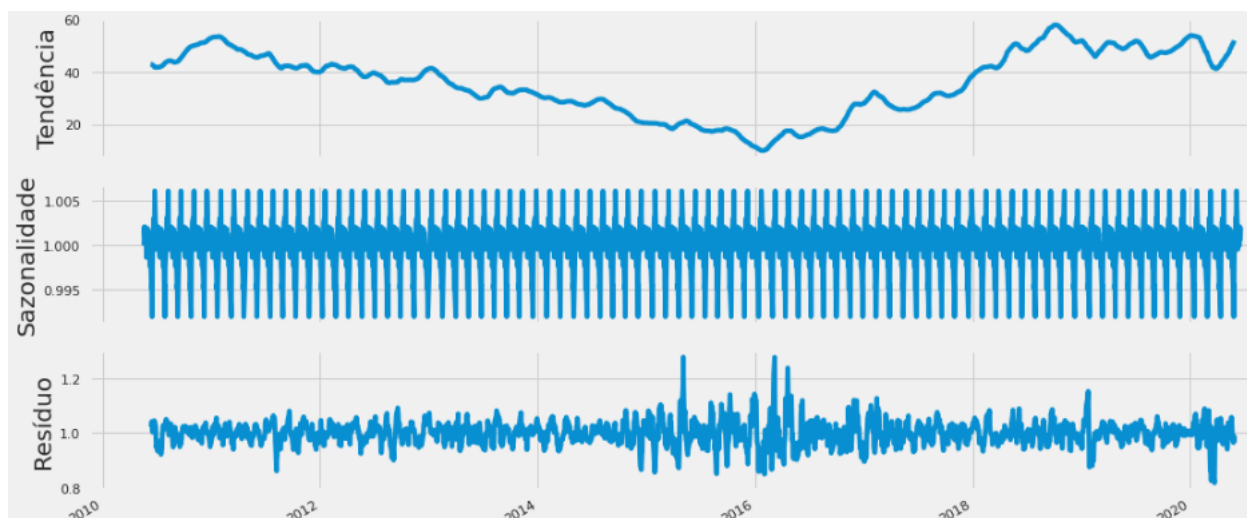


Figura 11. Decomposição da ST²¹.

A decomposição da ST demonstra que:

- Ocorreram 2 grandes tendências e nos últimos anos o preço não apresenta uma tendência evidente.
- A sazonalidade ocorre de 30 em 30 dias.
- De 2018 em diante os resíduos aumentaram. Neste mesmo período a ST não apresentou uma tendência clara.

4.2.3 Autocorrelação

Conforme mencionado em [Cowpertwait; Metcalfe \(2009\)](#), modelos autorregressivos como ARIMA, supõem que as observações anteriores são úteis para prever o valor da próxima etapa de tempo. Quanto mais forte a correlação entre a variável de saída e uma variável defasada específica, mais peso o modelo de auto-regressão pode colocar nessa variável. Esta característica se torna fundamental para conseguir comprovar que uma ST pode ser ou não prevista, [Parmezan \(2016\)](#). A tabela 08 mostra a autocorrelação do preço de VALE3 com diferentes atrasos.

Atraso em dias	Autocorrelação
----------------	----------------

²¹ Figura gerada pelo próprio autor.

1	0,997
30	0,945
313	0,501
314	0,499

Tabela 08. Resultado da autocorrelação dos dados do preço de fechamento das ações da VALE3.

Com estas informações é possível notar que até 313 dias ainda existe autocorrelação do preço. Na Figura 12 é mostrado a autocorrelação com atraso de 313 dias.

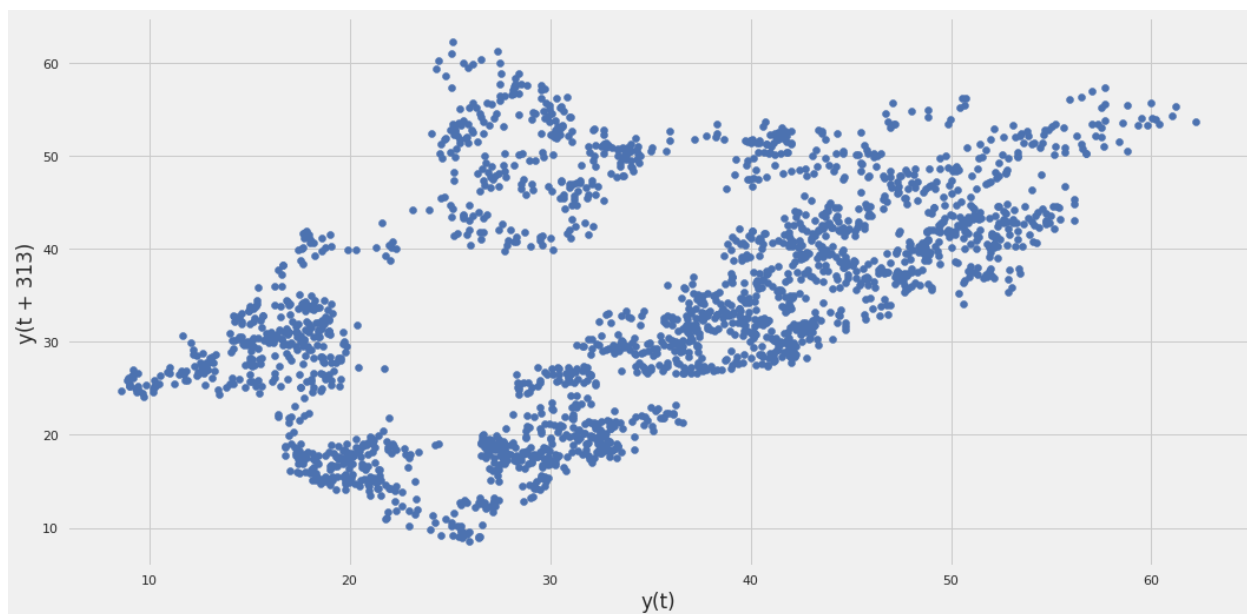


Figura 12. Autocorrelação do preço de fechamento com um atraso de 313 dias²².

É possível notar que há um grande grupo de observações ao longo de uma linha diagonal, mostrando claramente uma relação de correlação.

Para entender como a autocorrelação se comporta ao longo do tempo foi gerada a Figura 13 abaixo, tendo como intervalos de confiança 95% e 99%, conforme sugere a biblioteca de Pandas.

²² Figura gerada pelo próprio autor.

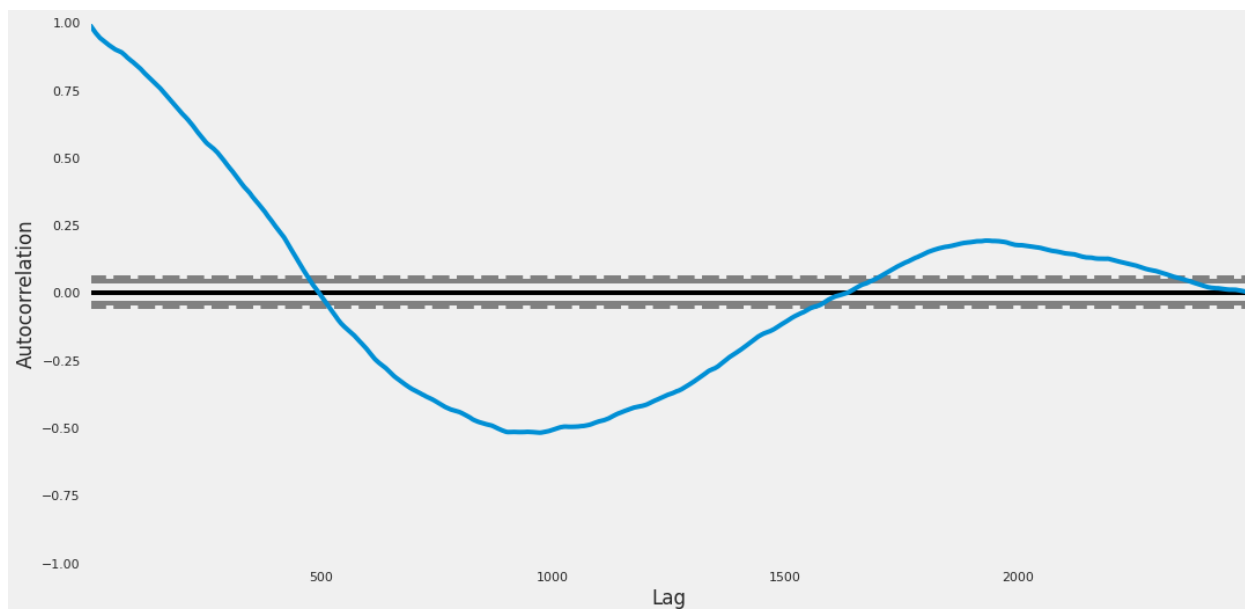


Figura 13. Autocorrelação do preço de fechamento da VALE3²³.

A partir da figura 13, é possível notar que:

- As auto correlações da ST analisada decaem com o passar do tempo.
- Os intervalos de confiança de 95% e 99% formam uma região onde os valores de autocorrelação fora dessa região são muito provavelmente uma correlação e não um acaso estatístico.

4.3 Divisão e Treinamento dos Dados

No processo de previsão de valores de uma ST demonstrado em [Parmezan \(2016\)](#), a primeira etapa consiste em dividir os dados em dois grupos, um para treinamento dos modelos e outro para testes. Neste trabalho foi realizada a separação dos dados sendo 2415 linhas para treinamento e 313 linhas para testes. O fato de escolher 313 é porque este é o limiar de autocorrelação do preço.

²³ Figura gerada pelo próprio autor.

As figuras 14 e 15 mostram como ficou a distribuição dos dados após a separação.

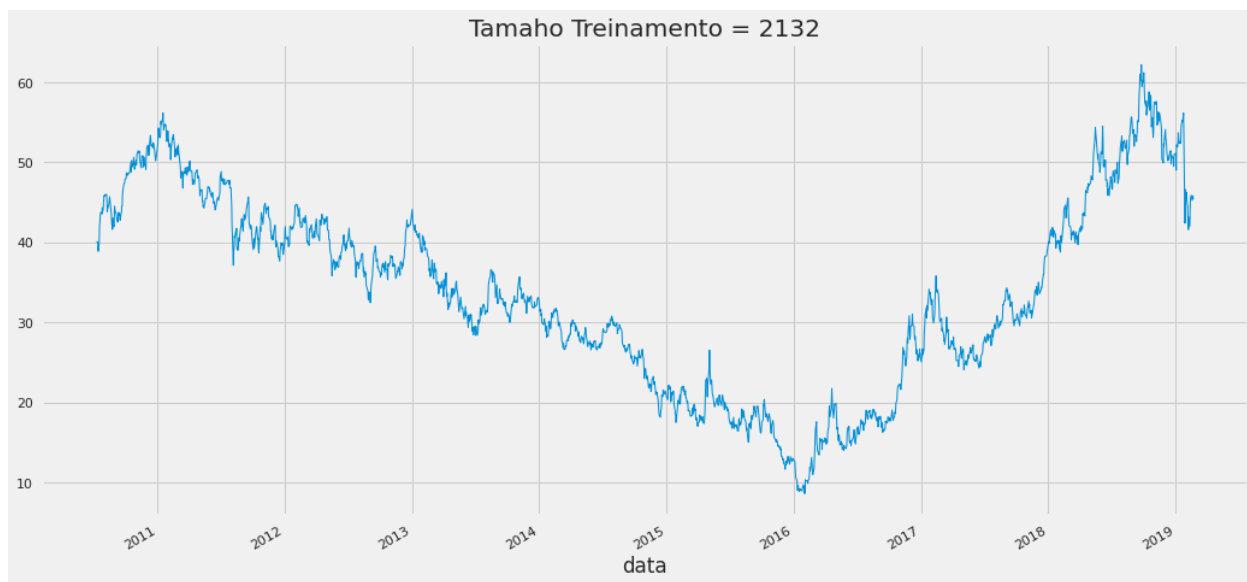


Figura 14. Distribuição dos dados de treinamento²⁴.

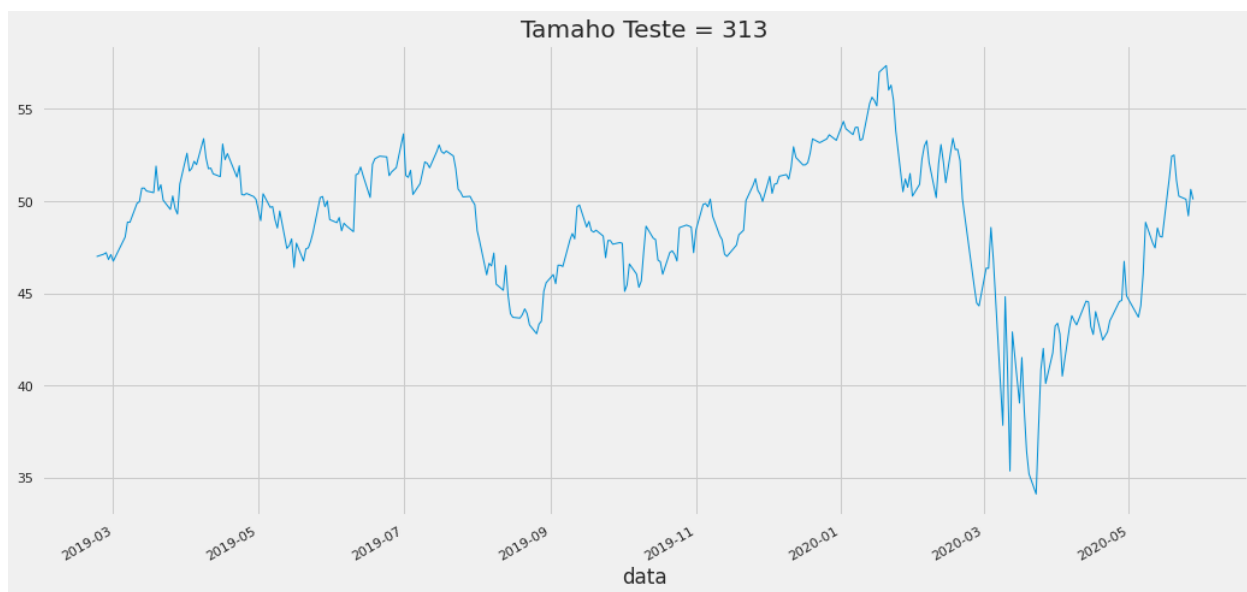


Figura 15. Distribuição dos dados de teste²⁵.

²⁴ Figura gerada pelo próprio autor.

²⁵ Figura gerada pelo próprio autor.

4.4 Implementação dos Previsores Estatísticos

Conforme mencionado no início deste capítulo, este trabalho seguiu a metodologia de [BOX et al, \(2015\)](#), onde na primeira etapa foi realizada a exploração dos dados para identificar as características da ST.

A identificação dos valores de p , d e q , do ARIMA e SARIMA são obtidas na aplicação de critérios de informação. Nesse sentido, o Critério de Informação de Akaike (AIC), expresso pela equação 3 ([Cowpertwait; Metcalfe, 2009](#)), penaliza a qualidade do ajuste do modelo.

$$AIC = -2 \times LL + (\log(n) + 1) \times NP \quad 3$$

4.4.1 Modelo AR

O modelo AR recebeu como parâmetro um atraso de 313 dias e foi removida a tendência e sazonalidade dos dados. O MAPE obtido foi de 7,96 % e o MSE foi de 21.78. A figura 16 compara os dados de teste e os dados previstos.

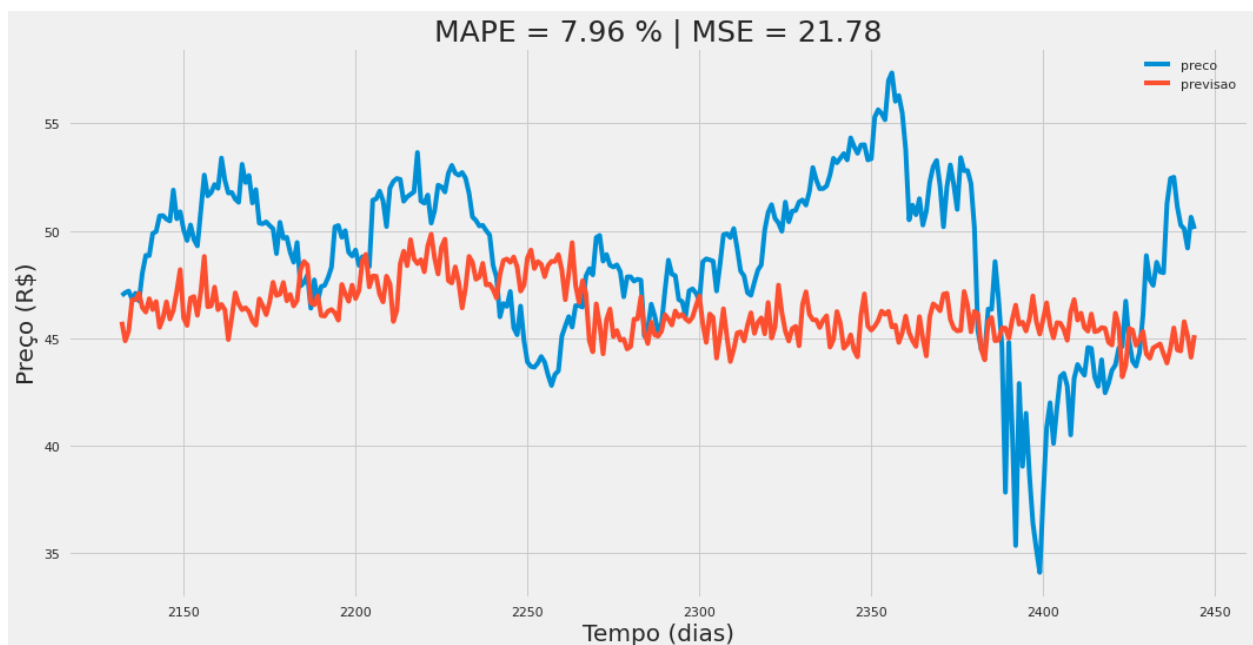


Figura 16. Comparação entre dados de testes e dados previstos pelo modelo AR.

4.4.1.1 Análise dos Resultados

O modelo AR apresenta uma tendência linear e não conseguiu capturar pontos relevantes de alta e baixa. Este modelo não é adequado para a previsão de ST financeiras, devido a falta de tendência e por não capturar períodos sazonais.

4.4.2 Modelo ARIMA

O modelo ARIMA recebeu como parâmetros (p, d, q) os valores de $(2, 1, 2)$ calculados com a função 03. Foi removida a tendência e sazonalidade dos dados. O MAPE obtido foi de 8,51 % e o MSE foi de 24,22. A figura 17 compara os dados de teste e os dados previstos.

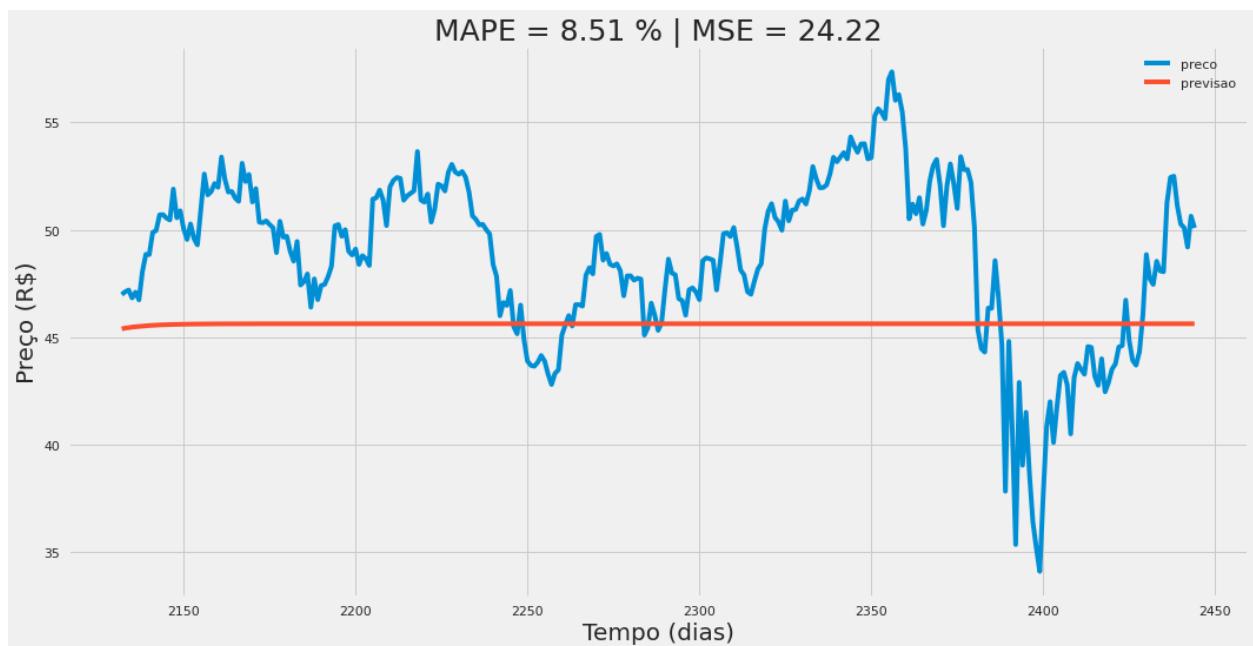


Figura 17. Comparação entre dados de testes e dados previstos pelo modelo ARIMA.

4.4.2.1 Análise dos Resultados

Assim como o modelo AR, o ARIMA apresenta um valor de MAPE próximo de 8%. É possível notar ainda que o modelo segue um padrão quase linear. Para dados de ST estocásticas

este modelo é inviável. Este resultado se deve ao fato que modelos ARIMA se comportam melhor com ST estacionárias, o que não é o caso deste conjunto de dados.

4.4.3 SARIMA

O modelo SARIMA recebeu como parâmetros (p, d, q) os valores de (2, 1, 2) calculados com a função 03. Foi removida a tendência e mantida a sazonalidade dos dados. O MAPE obtido foi de 8,42 % e o MSE foi de 23,67. A figura 18 compara os dados de teste e os dados previstos.

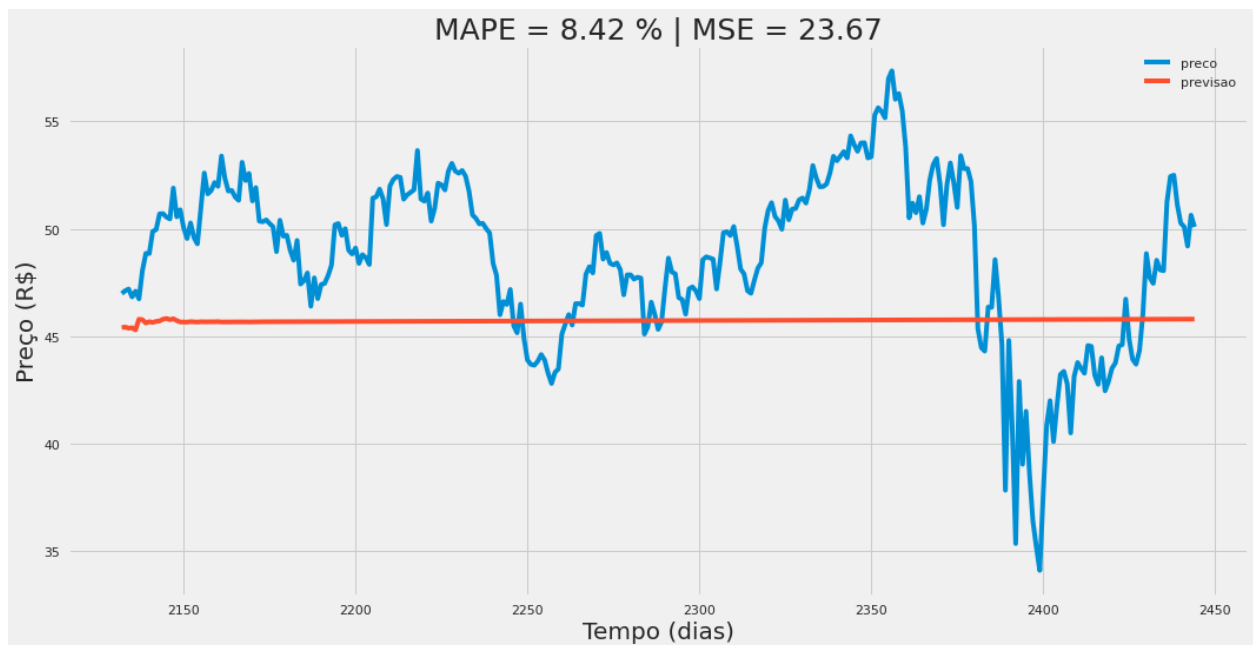


Figura 18. Comparação entre dados de testes e dados previstos pelo modelo SARIMA.

4.4.3.1 Análise dos Resultados

Este é o modelo econométrico que mais trabalha com características da ST. Apesar disso, não identificou a tendência correta, não capturou as oscilações, nem identificou a sazonalidade. Este resultado, assim como no caso do AR e ARIMA, se deve ao fato que modelos ARIMA e suas derivações se comportam melhor com ST estacionárias.

4.5 Implementação dos Previsores de AM

Para garantir a comparação entre os modelos econométricos e de AM foi necessário trabalhar somente com uma variável, haja vista que os modelos ARIMA são univariados, [Manav Kaushik et al, \(2020\)](#). Partindo deste pressuposto, foram geradas novas características que serviram como entrada nos modelos de Floresta Aleatória e SVR. As novas características são as seguintes:

- Resíduos: dados obtidos através da decomposição da ST.
- Tendência: dados obtidos através da decomposição da ST.
- Sazonalidade: dados obtidos através da decomposição da ST.
- Diferença 1, 2, 3, 4 e 5: dados obtidos calculando a diferença do preço do dia atual menos o preço de um, dois, três, quatro e cinco dias atrás.

4.5.1 Floresta Aleatória

O modelo de Floresta Aleatória recebeu como parâmetros, 1000 estimadores e árvores com profundidade máxima de 5 níveis.

O MAPE obtido foi de 2,90% e o MSE foi de 3,26. A figura 19 compara os dados de teste e os dados previstos.

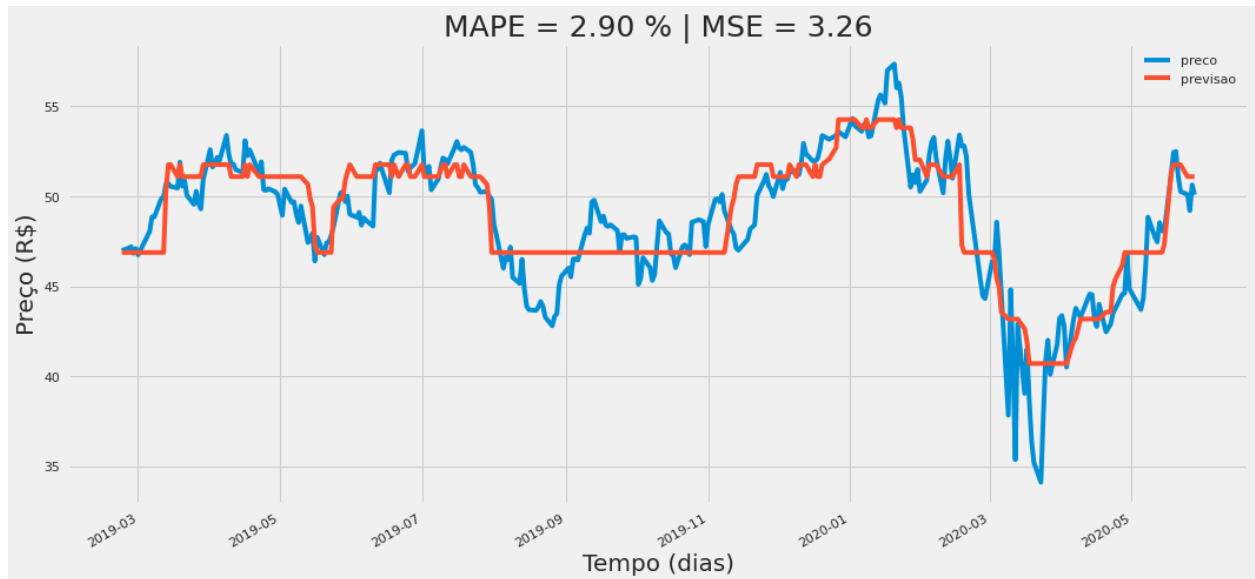


Figura 19. Comparação entre dados de testes e dados previstos pelo modelo de Floresta Aleatória.

4.5.1.1 Análise dos Resultados

As florestas aleatórias para a previsão de ST financeiras são uma boa opção para previsão de preço devido ao fato de que são fáceis de interpretar e entender os resultados. O modelo de floresta aleatória apresentado acima conseguiu identificar a tendência do movimento, porém não conseguiu definir as pequenas oscilações.

4.5.2 SVR

O modelo de SVR foi configurado com o kernel radial e gamma de 0,1 para garantir uma boa generalização do modelo, conforme utilizado em [Patel et al, \(2015\)](#).

O MAPE obtido foi de 4,41 % e o MSE foi de 10,13. A figura 20 compara os dados de teste e os dados previstos.

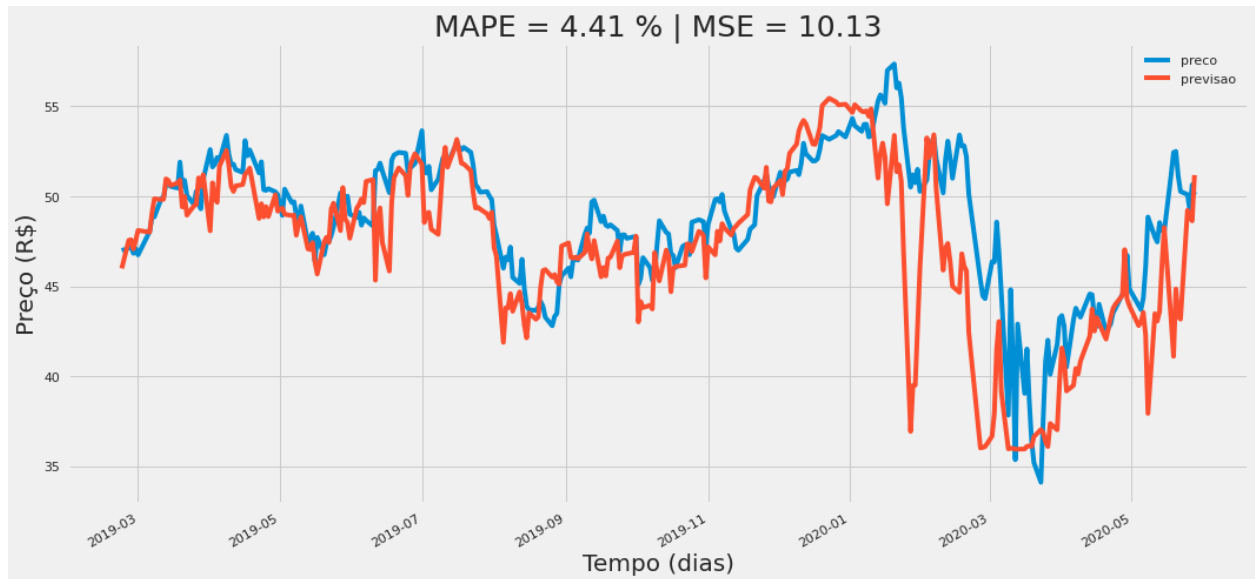


Figura 20. Comparação entre dados de testes e dados previstos pelo modelo SVR.

4.5.2.1 Análise dos Resultados

Ao contrário da Floresta Aleatória, este modelo conseguiu prever oscilações de baixa e alta com mais precisão. O problema apresentado é na amplitude do movimento da ST, no qual a SVR consegue identificar que houve alta ou baixa mas acaba prevendo movimentos mais amplos do que o esperado em uma tendência de baixa e prevendo movimentos mais suaves numa tendência de alta.

4.6 Implementação dos Previsores de AP

O único modelo de AP utilizado neste trabalho foi uma LSTM. A RNA possui a seguinte configuração:

- Camada de entrada: 50 unidades de LSTM e regularização dos pesos em 0,2
- Camadas ocultas:
 - 1ª: 10 unidades de LSTM e regularização dos pesos em 0,2
 - 2ª: 10 unidades de LSTM e regularização dos pesos em 0,2

- 3ª: 10 unidades de LSTM e regularização dos pesos em 0,2
- Camada de saída: 1 unidade densa

O treinamento da LSTM foi feito utilizando lotes de 30 em 30 linhas e foram percorridas 1000 épocas.

O MAPE obtido foi de 2,78 % e o MSE foi de 3,23. A figura 21 compara os dados de teste e os dados previstos.

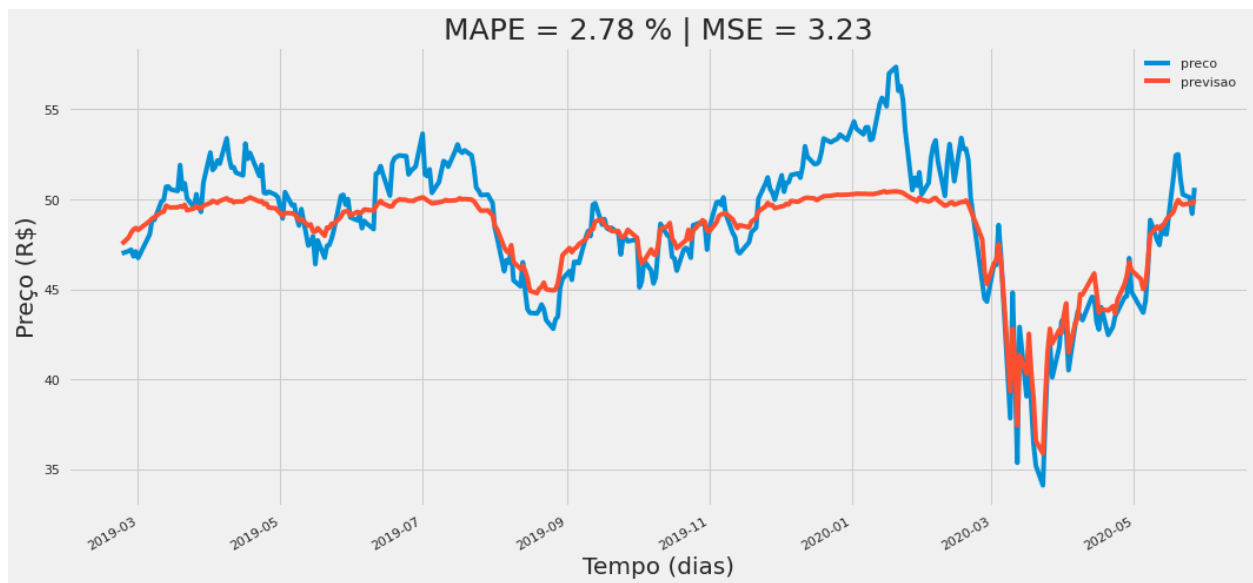


Figura 21. Comparação entre dados de testes e dados previstos pelo modelo de LSTM.

4.6.1 Análise dos Resultados

Modelos de múltiplas camadas, como neste caso, enfrentam o problema de overfitting e redução de erro de retropropagação, além de problemas em determinar automaticamente as defasagens de tempo ideais durante o ajuste dados de ST, [Manav Kaushik et al, \(2020\)](#).

Para entender se o modelo em questão sofreu algum problema de ajuste é necessário analisar a taxa de perda durante a fase de treinamento e validação. Na figura 22 há um gráfico onde a taxa de perda (loss) dos dados de treinamento e validação diminuem e se estabilizam em torno do mesmo ponto.

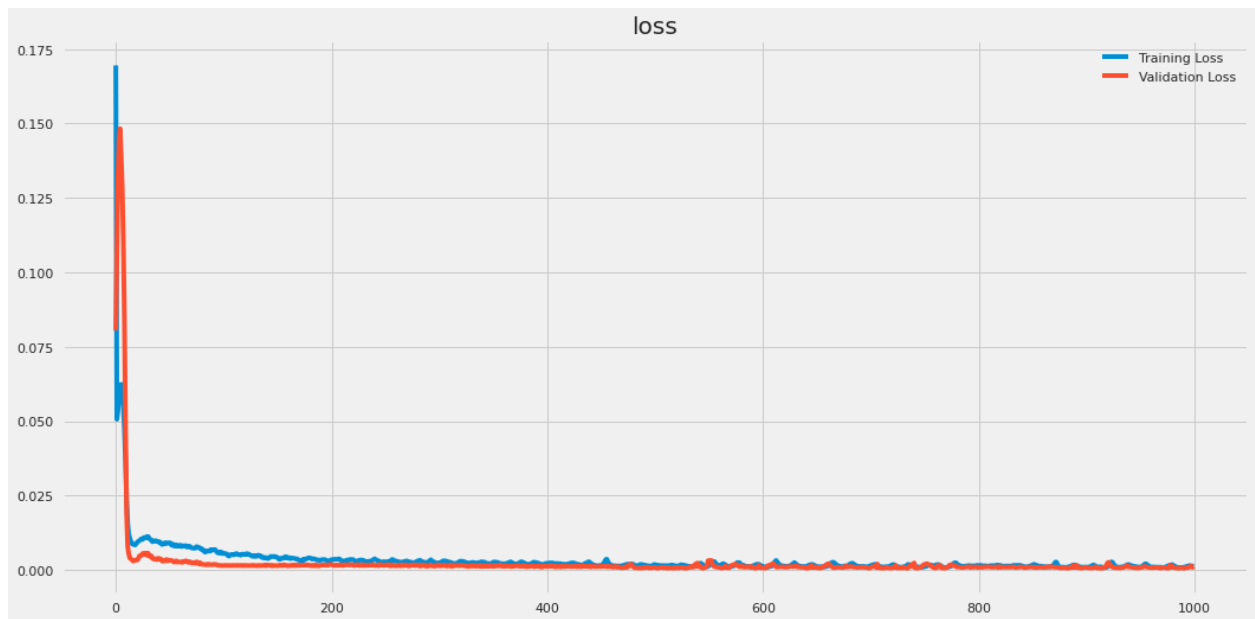


Figura 22. Convergência da taxa de perda entre os dados de treinamento e validação.

O modelo convergiu a taxa de perda, além disso, o modelo não sofreu problema de sobreajuste pois os valores de perda (loss) de treinamento e validação estão muito próximos após a convergência.

Este foi o modelo que apresentou o melhor MSE. Conforme mencionado em [Manav Kaushik et al, \(2020\)](#), modelos de LSTM podem sofrer com problemas de ajuste. Devido a isso foi necessário validar o treinamento separando os dados em mais um conjunto, o de validação. Após a confirmação que o modelo não apresentou os problemas de ajuste é possível dizer que é o melhor modelo, dentre todos os comparados, para prever ST no contexto de mercados financeiros.

4.7 Considerações Finais

Neste Capítulo foi descrito o processo de análise dos dados, desde a fase de obtenção, manipulação, pré-processamento, análise exploratória da ST, divisão e treinamento até a fase de implementação dos previsores e análise dos resultados de cada modelagem.

Para o desenvolvimento das técnicas econométricas foi utilizada a metodologia de [BOX et al, \(2015\)](#) e para as modelagens de AM e AP foi aplicada a metodologia demonstrada em [Parmezan \(2016\)](#).

Os resultados obtidos são comparados na tabela 09.

Modelo / Performance	AR	ARIMA	SARIMA	Floresta Aleatória	SVM	LSTM (RNA)
MAPE	7,95 %	8,51 %	9,53 %	2,90 %	4,41 %	2,78 %
RSE	21,78	24,22	33,96	3,26	10,13	3,23

Tabela 09. Comparação dos resultados de MAPE e RSE de cada método de predição²⁶

Os modelos econométricos, AR, ARIMA e SARIMA apresentaram desempenho inferior quando comparado com os modelos de aprendizagem, Floresta Aleatória, SVM e LSTM.

Já os modelos de aprendizagem obtiveram resultados superiores. Este resultado confirma o que foi discutido em 2.3, no qual diz que os métodos de AM e AP, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos, sendo assim não precisam supor a linearidade e estacionariedade dos dados.

²⁶ Fonte: [Manav Kaushik et al, \(2020\)](#)

5 CONCLUSÕES E ATIVIDADES FUTURAS

Neste trabalho foi apresentado uma comparação dos métodos mais utilizados na literatura para a previsão de ST estocásticas. Foi realizado também um processo completo de mineração de dados seguindo metodologias tradicionais.

A análise qualitativa da literatura demonstrou que a previsão de ST se adaptou ao longo do tempo devido a criação e aperfeiçoamento de técnicas de AM e AP. As publicações passaram a adotar cada vez mais modelos de aprendizagem em razão dos bons resultados. Contudo, os modelos econométricos não foram descartados das pesquisas pois ainda apresentaram vantagens, como por exemplo, a facilidade em descrever o comportamento da ST.

Além do entendimento do estado da arte na predição de ST, este trabalho teve como objetivos específicos a obtenção, análise, manipulação, processamento, exploração estatística, divisão, treinamento, backtesting e avaliação dos resultados das modelagens econométricas e de aprendizagem.

Para cada modelagem foi realizada uma análise dos resultados utilizando as medidas de MAPE e MSE e elaborado uma tabela comparativa conforme previsto no objetivo 6.

O modelo de LSTM obteve o melhor resultado de MSE e de MAPE. Já os modelos econométricos, AR, ARIMA e SARIMA, para o dados de ST com um padrão estocástico não se saíram bem devido ao fato de que a ST a ser prevista não apresenta um comportamento estacionário. Esta conclusão vai de encontro com o que foi mencionado em (Bueno, 2011), onde diz que as inferências estatísticas só terão validade se os resíduos da ST estimada forem estacionários, o que não é o caso de índices e preço de ações.

Após a realização do trabalho e a avaliação dos resultados obtidos, é possível concluir que os objetivos específicos levantados foram alcançados.

Durante a execução do projeto foi percebido oportunidades de possíveis trabalhos futuros, conforme citado abaixo:

- Comparar técnicas econométricas, como ARCH, GARCH e VAR.
- Comparar técnicas de AM, como algoritmos de Bayes e Ensembles.

- Comparar técnicas de Aprendizado por Reforço.
- Para todos os modelos analisados poderiam ser analisados considerando mais atributos, como abertura, fechamento, máximo e mínima do preço, volume e até dados que envolvem uma análise mais fundamentalista, como balanço, dívida da empresa, taxa de juros, entre outras informações.

REFERÊNCIAS

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015) “**Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques**”, Expert Systems with Applications, 42, 259-268. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417414004473>

Sampaio e Mancini (2007) “**Estudos De Revisão Sistemática: Um Guia Para Síntese Criteriosa Da Evidência Científica**”, Universidade Federal de Minas Gerais, Belo Horizonte, MG. Disponível em : <https://www.scielo.br/pdf/rbfis/v11n1/12.pdf>

Lo, Andrew W. (2004) “**The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective**”, Journal of Portfolio Management, Forthcoming. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=602222

Kaushik, Manav, e A. K. Giri. (2020) “**Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques**”, arXiv preprint arXiv:2002.10247. Disponível em: <https://arxiv.org/abs/2002.10247>

Luger, George F. (2013) **Inteligência Artificial**. 6 ed. São Paulo, Brasil. Pearson Education.

BUSSAB, Wilton de Oliveira e MORETTIN, Pedro Alberto (2010) **Estatística básica**. 6 ed. São Paulo, Brasil. Editora Saraiva.

Cao, L., Tay, F. (2001) “**Financial Forecasting Using Support Vector Machines**”, *Neural Comput & Applic* 10, 184–192. Disponível em: <https://link.springer.com/article/10.1007/s005210170010>

Parmezan, Antonio Rafael Sabino (2016) “**Predição de séries temporais por similaridade**”, Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2016. doi:10.11606/D.55.2016. tde-21112016-150659. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-21112016-150659/pt-br.php>

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M (2015) **Time series analysis: Forecasting and control**. 5. ed. New Jersey, United States of America: Wiley. (Wiley Series in Probability and Statistics).

Junior, Salomon and Pamplona (2014) “**ARIMA: An Applied Time Series Forecasting Model for the Bovespa Stock Index**” *Applied Mathematics*, 5, 3383-3391 Published Online December 2014 in SciRes. Disponível em: <https://www.researchgate.net/publication/275214476>

Kaastra, M. Boyd (1996) “**Designing a neural network for forecasting financial and economic time series**”, *Neurocomputing*, 10 (3), pp. 215-236. 1996 Disponível em: <https://www.sciencedirect.com/science/article/pii/0925231295000399/pdf?md5=85cce437dd80a9cc55ebaa8a43455327&pid=1-s2.0-0925231295000399-main.pdf>

Daniel, Fabrice (2019) “**Financial Time Series Data Processing for Machine Learning**”, arXiv preprint arXiv:1907.03010, 2019 - arxiv.org. Disponível em: <https://arxiv.org/pdf/1907.03010.pdf>

Sergio, Anderson Tenório (2017) “**Seleção dinâmica de combinadores de previsão de séries temporais**”, Dissertação de Doutorado em Ciências de Computação - Centro de Informática, Universidade Federal de Pernambuco, Recife. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/25449/1/TESE%20Anderson%20Ten%C3%B3rio%20Sergio.pdf>

Cowpertwait, P. S. P. Metcalfe, A. V. (2009) “**Introductory time series with R**”, Springer Science & Business Media.

Vapnik, V. N. (1999) “**An overview of statistical learning theory**”, IEEE Transactions on Neural Networks, 10, 988–999. Disponível em : <https://ieeexplore.ieee.org/document/788640>

PYTHON SOFTWARE FOUNDATION (2021) **Python Language Site: Documentation**. Página de documentação. Disponível em: <https://www.python.org/doc/>

GIL, A. C. (2010) **Como elaborar projetos de pesquisa**, 4 ed. São Paulo, Brasil. Editora Atlas.

Bueno, R. L. (2011) **Econometria de Séries Temporais**. 2 ed. São Paulo, Brasil. Editora Cengage Learning.

APÊNDICE A: Artigo

Análise Comparativa de Técnicas para a Previsão de Séries Temporais no Contexto de Mercados Financeiros

Bruno A. R. M. Campos¹, Dr. Élder Rizzon Santos¹, Dra. Andrea Cristina Konrath¹

¹Departamento de informática e Estatística - Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 - 88040-900 - Florianópolis - SC - Brazil

brunocampos01@gmail.com, {elder.santos, andrea.ck}@ufsc.br

Abstract. *To produce the comparison of the main prediction techniques for Time Series in the financial market context, qualitative analyses were performed in order to understand the state of the art on Time Series prediction and forecasting theories in financial markets. In addition, sourcing, preparation and modeling processes were performed to ensure a standardization of the input data in each model used. Finally, a comparative analysis of the predictors' results was performed.*

Resumo. *Para produzir a comparação das principais técnicas de previsão de Séries Temporais no contexto de mercado financeiro foi realizada uma análise qualitativa a fim de compreender o estado da arte sobre previsão de Séries Temporais e teorias de previsão nos mercados financeiros. Além disso, foram realizados processos de obtenção, preparação e modelagem para garantir uma padronização dos dados de entrada em cada modelo utilizado. Por fim, foi realizada uma análise comparativa dos resultados dos preditores.*

1. Introdução

Prever valores resultantes no tempo não é uma tarefa simples, ainda mais no contexto de mercados financeiros. A razão se deve às Séries Temporais (ST) financeiras que por natureza são estocásticas, ruidosas, e não estacionárias como explicado por Manav Kaushik et al, (2020). Estas características corroboram a Hipótese de Mercados Adaptativos (HMA) proposta por Lo (2004), que trata da conciliação da Hipótese de Eficiência de Mercados (HEM) com a economia comportamental, onde a oscilação de valor de uma ação pode ser afetada não somente pelo comportamento do histórico do preço mas também por informações macroeconômicas e sentimentos de investidores, tais como aversão à perda e excesso de confiança. Sabendo disso, é possível observar ao longo do tempo que determinados padrões de tomada de decisão dos investidores tornam-se recorrentes, isso faz com que sejam criadas várias estratégias para maximizar os lucros em negociações.

Segundo Lo (2004), como nem toda informação financeira é explícita no mercado, ineficiências acabam afetando as estratégias de negociação. É neste momento que o uso de métodos computacionais pode se tornar uma vantagem competitiva. O mercado financeiro é constantemente estudado utilizando a Inteligência Artificial (IA) junto com a estatística.

A IA pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente (Luger, 2013). Dentro da IA existe o campo de estudos de Aprendizado de Máquina (AM) que faz uso de algoritmos e métodos que possibilitam computadores a aprender por experiências e dados analíticos (Luger, 2013).

Em cenários onde o volume, velocidade e variedade dos dados são fatores determinantes para o lucro, se torna inviável fazer análises somente com indicadores estatísticos, é então neste momento que se faz necessário o uso do AM. Além desse motivo, há fatores do comportamento humano, conforme mencionado na HMA que afetam a variação de preço. Por estes aspectos, utilizar meios de processar, reconhecer padrões e testar estratégias pode ser fundamental para negociações de sucesso.

O AM e o aprendizado profundo (AP) são subáreas da Inteligência Artificial que oferecem suporte para resolver problemas como a previsão de ST financeiras. Devido a isto, várias técnicas de AM vêm sendo testadas para melhor prever as ST como por exemplo, máquinas de vetor de suporte, floresta aleatória, redes bayesianas e até técnicas de AP, como Redes Neurais Artificiais (Patel et al, 2015).

Embora mais antigos, modelos estatísticos comumente utilizados na área econométrica, ainda são estudados e utilizados com resultados satisfatórios para prever ST. Entre os principais preditores, podemos citar modelos lineares como AR (Autoregressive), MA (Moving Average) e ARIMA (Autoregressive Integrated Moving Average) e não-lineares como GARCH (Generalized ARCH). Partindo disso, as técnicas econométricas tradicionais, também chamadas de técnicas estatísticas, estabelecem referências econômicas devido à sua ampla utilização na literatura (Manav Kaushik et al, 2020). Mesmo assim, cabe ressaltar que estes modelos apresentam limitações devido a duas hipóteses que sempre os acompanham: (i) os dados a serem preditos são lineares e (ii) a ST é estacionária (Cao e Tay, 2001).

Este artigo visa analisar as principais modelagens matemáticas para a previsão de ST financeiras. O intuito é demonstrar com experiência prática que os modelos de AM e AP proporcionam resultados mais competitivos do que os modelos econométricos.

2. Fundamentação Teórica

2.1. Séries Temporais

Segundo Sérgio (2017) uma ST pode ser definida como um conjunto de observações coletadas ao longo do tempo, em uma determinada ordem. Do ponto de vista estatístico, uma série de dados históricos pode ser tratada como uma sequência de variáveis aleatórias. Uma ST pode então ser referida como um processo estocástico discreto ao longo do tempo.

Cada dado coletado pode ser um ponto em um gráfico de duas dimensões, onde o eixo das ordenadas determina as medições dos dados e o eixo das abscissas delimita em que momento discreto do tempo tais medições foram aferidas.

2.2 Previsores Estatísticos

O modelo de BOX et al. (2015), também conhecido como auto-regressivo integrado de média móvel, ARIMA (AutoRegressive Integrated Moving Average), foi elaborado a partir da coletânea de diversos trabalhos publicados na época, trazendo como mérito, a unificação das principais técnicas estatísticas, em uma única metodologia para construir modelos de previsão de ST.

Conforme mencionado em Parmezan (2016), a seleção de modelos ARIMA é apropriada quando a série em estudo é estacionária, ou seja, suas propriedades estatísticas básicas, como média, variância e covariância, permanecem constantes ao longo do tempo.

De acordo com BOX et al. (2015), qualquer processo estocástico pode ser representado por um modelo auto-regressivo ou médias móveis. No caso de processos estocásticos não estacionários, aplica-se o operador de diferenciação para tornar a série estacionária.

2.3 Previsores de Aprendizado de Máquina

De acordo com Parmezan (2016), os métodos de AM para previsão, em oposição aos modelos estatísticos, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos. Devido a isso, são mais simples de serem ajustados e alcançam considerável desempenho mesmo quando aplicados à ST complexas e altamente não-lineares. Modelos de AM são orientados a dados, o que significa dizer que, a princípio, não há necessidade de entender o conjunto de dados no qual o algoritmo será treinado. Por exemplo, dada uma ST, não é preciso decidir de antemão a respeito da linearidade ou não do modelo pois os algoritmos são responsáveis pelo ajuste automático. Essa característica é uma vantagem do AM frente a modelos estatísticos.

2.4 Previsores de Aprendizado Profundo

As Redes Neurais Artificiais (RNA) representam uma abordagem em inteligência artificial (IA) e suas constituições simulam a estrutura de funcionamento dos neurônios no cérebro humano (Luger, 2013). Uma RNA, similarmente ao seu paralelo biológico, é composta por determinado número de neurônios conectados por conexões sinápticas. Cada sinapse possui um peso atribuído, onde o conhecimento acumulado na rede depende, diretamente, da atualização desses pesos. Cada neurônio, ou também chamado de perceptron, possui um conjunto de entradas e de saídas ligadas a outros neurônios. Um neurônio possui um patamar de ativação que deriva da conjugação das entradas atuais recebidas com os pesos sinápticos das conexões, cujo valor resultante poderá ativar ou não a saída desse neurônio, (Luger, 2013).

3. Desenvolvimento

O MetaTrader²⁷ é uma plataforma open source para negociação de ativos financeiros. A obtenção dos dados foi feita através da plataforma, a qual permitiu baixar uma tabela de dados no formato csv.

O período escolhido para a modelagem foi do dia 20 de maio de 2010 até 19 de junho de 2020, sendo uma granularidade diária de informação, ou seja, para cada dia foi obtido uma linha contendo as informações da data e preço de fechamento. O total de linhas obtidas foi de 2495.

Para explorar os dados foram analisadas as medidas de posição. Conforme mencionado em Morettin e Bussab (2010), as medidas de posição servem para resumir os dados através de um ou mais valores representativos para a ST. A tabela 1 traz as medidas calculadas para o preço de fechamento das ações da VALE3.

Tabela 1. Medidas de posição dos dados do preço das ações da VALE3.

Medidas	Preço R\$
média	35,80
mediana	36,88
moda	27,43
mínimo	8,60
máximo	62,20

3.1 Implementação dos Previsores Estatísticos

3.1.1 Modelo AR

O modelo AR recebeu como parâmetro um atraso de 313 dias e foi removida a tendência e sazonalidade dos dados. O MAPE obtido foi de 7,96 % e o MSE foi de 21.78. A figura 1 compara os dados de teste e os dados previstos.

²⁷ Endereço eletrônico da plataforma: <https://www.metatrader5.com/pt>

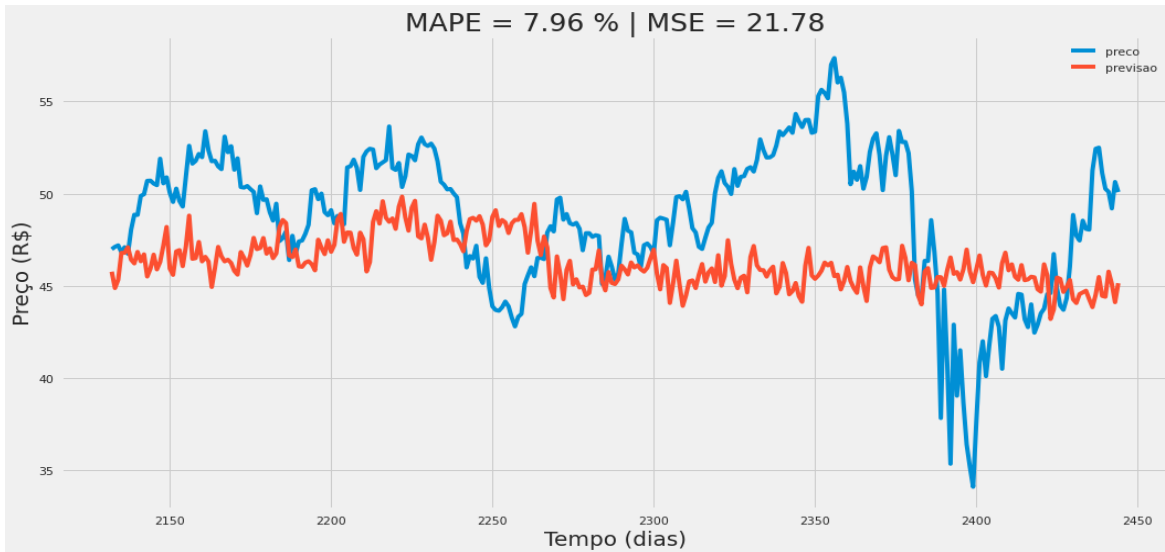


Figura 1. Comparação entre dados de testes e dados previstos pelo modelo AR.

O modelo AR apresenta uma tendência linear e não conseguiu capturar pontos relevantes de alta e baixa. Este modelo não é adequado para a previsão de ST financeiras, devido a falta de tendência e por não capturar períodos sazonais.

3.1.2 Modelo ARIMA

Para o modelo ARIMA foi removido a tendência e sazonalidade dos dados. O MAPE obtido foi de 8,51 % e o MSE foi de 24,22. A figura 2 compara os dados de teste e os dados previstos.

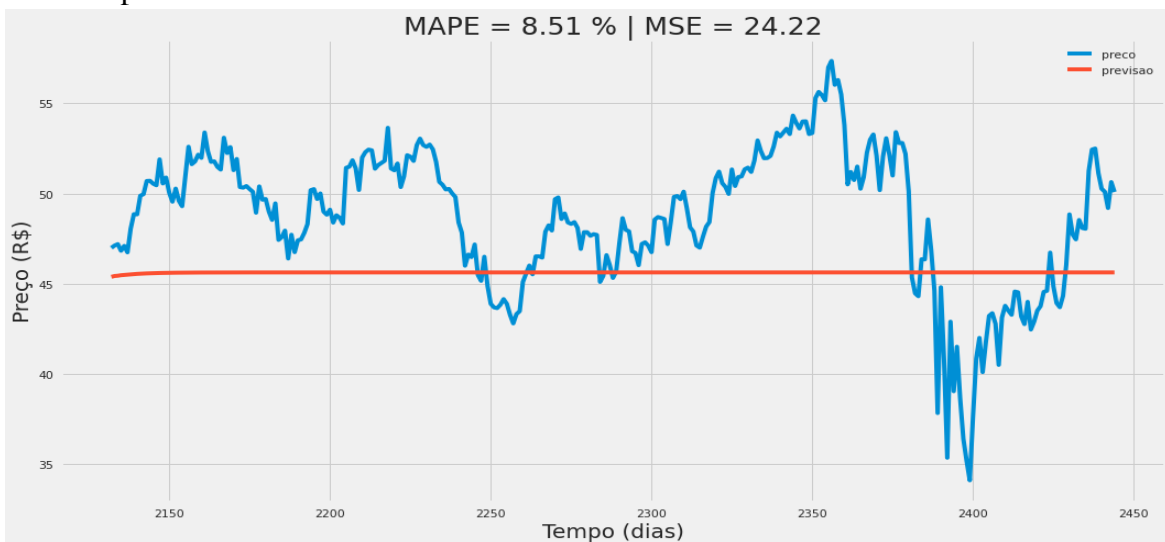


Figura 2. Comparação entre dados de testes e dados previstos pelo modelo ARIMA.

Assim como o modelo AR, o ARIMA apresenta um valor de MAPE próximo de 8%. É possível notar ainda que o modelo segue um padrão quase linear. Para dados de ST

estocásticas este modelo é inviável. Este resultado se deve ao fato que modelos ARIMA se comportam melhor com ST estacionárias, o que não é o caso deste conjunto de dados.

3.2 Implementação dos Previsores de AM

O modelo de SVR foi configurado com o kernel radial e gamma de 0,1 para garantir uma boa generalização do modelo, conforme utilizado em Patel et al, (2015).

O MAPE obtido foi de 4,41 % e o MSE foi de 10,13. A figura 3 compara os dados de teste e os dados previstos.

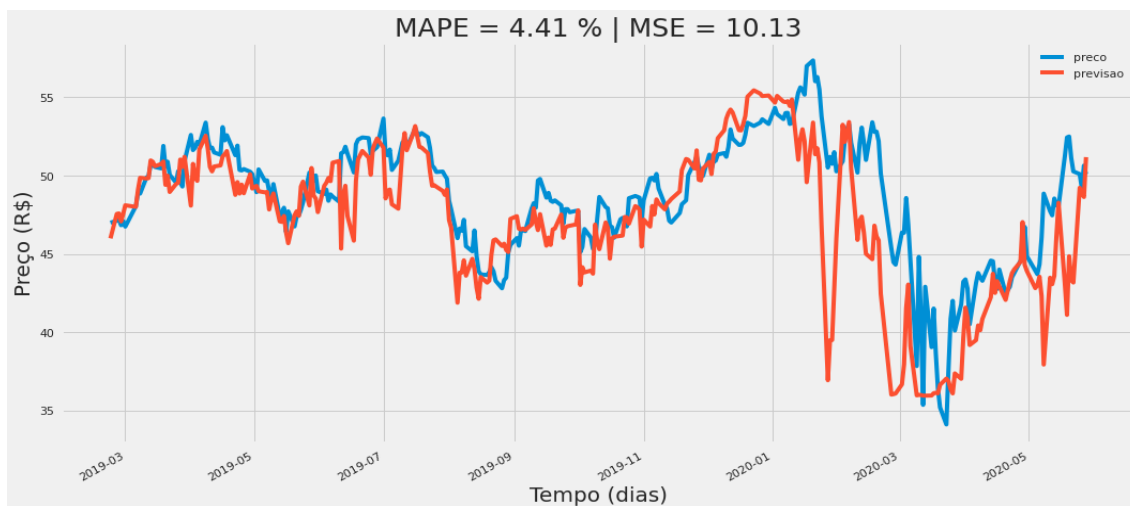


Figura 3. Comparação entre dados de testes e dados previstos pelo modelo SVR.

Este modelo conseguiu prever oscilações de baixa e alta com mais precisão. O problema apresentado é na amplitude do movimento da ST, no qual a SVR consegue identificar que houve alta ou baixa mas acaba prevendo movimentos mais amplos do que o esperado em uma tendência de baixa e prevendo movimentos mais suaves numa tendência de alta.

3.3 Implementação dos Previsores de AP

O único modelo de AP utilizado neste trabalho foi uma LSTM. A RNA possui a seguinte configuração:

- Camada de entrada: 50 unidades de LSTM e regularização dos pesos em 0,2
- Camadas ocultas:
 - 1^a: 10 unidades de LSTM e regularização dos pesos em 0,2
 - 2^a: 10 unidades de LSTM e regularização dos pesos em 0,2
 - 3^a: 10 unidades de LSTM e regularização dos pesos em 0,2
- Camada de saída: 1 unidade densa

O treinamento da LSTM foi feito utilizando lotes de 30 em 30 linhas e foram percorridas 1000 épocas.

O MAPE obtido foi de 2,78 % e o MSE foi de 3,23. A figura 4 compara os dados de teste e os dados previstos.

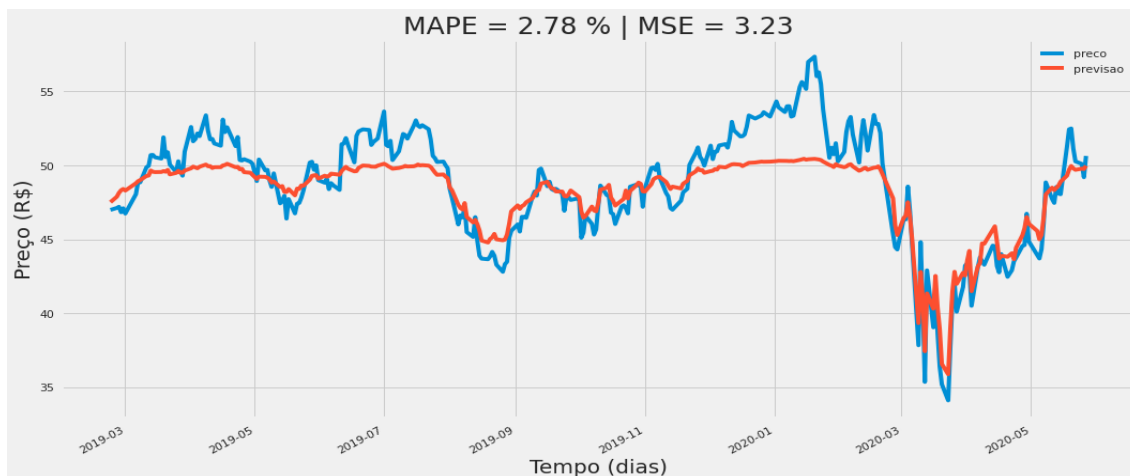


Figura 4. Comparação entre dados de testes e dados previstos pelo modelo de LSTM.

Modelos de múltiplas camadas, como neste caso, enfrentam o problema de overfitting e redução de erro de retropropagação, além de problemas em determinar automaticamente as defasagens de tempo ideais durante o ajuste dados de ST, Manav Kaushik et al, (2020).

Para entender se o modelo em questão sofreu algum problema de ajuste foi analisado a taxa de perda durante a fase de treinamento e validação. Como resultado o modelo convergiu a taxa de perda garantindo que não houve problemas de sobreajuste.

Este foi o modelo que apresentou o melhor MSE. Conforme mencionado em Manav Kaushik et al, (2020), modelos de LSTM podem sofrer com problemas de ajuste. Devido a isso foi necessário validar o treinamento separando os dados em mais um conjunto, o de validação. Após a confirmação que o modelo não apresentou os problemas de ajuste é possível dizer que é o melhor modelo, dentre todos os comparados, para prever ST no contexto de mercados financeiros.

3.4 Considerações Finais

Para o desenvolvimento das técnicas econométricas foi utilizada a metodologia de BOX et al, (2015) e para as modelagens de AM e AP foi aplicada a metodologia demonstrada em Parmezan (2016). Os resultados obtidos são comparados na tabela 5.

Tabela 2. Comparação dos resultados de MAPE e RSE de cada método de previsão.

Modelo / Performance	AR	ARIMA	SVM	LSTM (RNA)
MAPE	7,95 %	8,51 %	4,41 %	2,78 %
RSE	21,78	24,22	10,13	3,23

Os modelos econométricos, AR e ARIMA apresentaram desempenho inferior quando comparado com os modelos de aprendizagem, SVM e LSTM.

5. Conclusões e Atividades Futuras

Neste trabalho foi apresentado uma comparação dos métodos mais utilizados na literatura para a previsão de ST estocásticas. A análise qualitativa da literatura demonstrou que a previsão de ST se adaptou ao longo do tempo devido a criação e aperfeiçoamento de técnicas de AM e AP. As publicações passaram a adotar cada vez mais modelos de aprendizagem em razão dos bons resultados. Contudo, os modelos econométricos não foram descartados das pesquisas pois ainda apresentaram vantagens, como por exemplo, a facilidade em descrever o comportamento da ST.

O modelo de LSTM obteve o melhor resultado de MSE e de MAPE. Esta conclusão vai de encontro com o que foi mencionado em (Bueno, 2011), onde diz que as inferências estatísticas só terão validade se os resíduos da ST estimada forem estacionários, o que não é o caso de índices e preço de ações.

Já os modelos econométricos, AR e ARIMA para o dados de ST com um padrão estocástico não se saíram bem devido ao fato de que a ST a ser prevista não apresenta um comportamento estacionário.

Os métodos de AM e AP, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos, sendo assim não precisam supor a linearidade e estacionariedade dos dados.

Durante o desenvolvimento das predições foi percebido oportunidades de possíveis trabalhos futuros, conforme citado abaixo:

- Comparar técnicas econométricas, como ARCH, GARCH e VAR.
- Comparar técnicas de AM, como algoritmos de Bayes e Ensembles.
- Comparar técnicas de Aprendizado por Reforço.
- Para todos os modelos analisados poderiam ser analisados considerando mais atributos, como abertura, fechamento, máximo e mínima do preço, volume e até dados que envolvem uma análise mais fundamentalista, como balanço, dívida da empresa, taxa de juros, entre outras informações.

7. Referências

- Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015) “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques”, *Expert Systems with Applications*, 42, 259-268, <https://www.sciencedirect.com/science/article/pii/S0957417414004473>
- Sampaio e Mancini (2007) “Estudos De Revisão Sistemática: Um Guia Para Síntese Criteriosa Da Evidência Científica”, Universidade Federal de Minas Gerais, Belo Horizonte, MG, <https://www.scielo.br/pdf/rbfis/v11n1/12.pdf>
- Lo, Andrew W. (2004) “The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective”, *Journal of Portfolio Management*, Forthcoming, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=602222
- Kaushik, Manav, e A. K. Giri. (2020) “Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques”, arXiv preprint arXiv:2002.10247, <https://arxiv.org/abs/2002.10247>
- Luger, George F. (2013) *Inteligência Artificial*. 6 ed. São Paulo, Brasil. Pearson Education.
- BUSSAB, Wilton de Oliveira e MORETTIN, Pedro Alberto (2010) *Estatística básica*. 6 ed. São Paulo, Brasil. Editora Saraiva.
- Cao, L., Tay, F. (2001) “Financial Forecasting Using Support Vector Machines”, *Neural Comput & Applic* 10, 184–192, <https://link.springer.com/article/10.1007/s005210170010>
- Parmezan, Antonio Rafael Sabino (2016) “predição de séries temporais por similaridade”, Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2016. doi:10.11606/D.55.2016. tde-21112016-150659, <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-21112016-150659/pt-br.php>
- BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M (2015) *Time series analysis: Forecasting and control*. 5. ed. New Jersey, United States of America: Wiley. (Wiley Series in Probability and Statistics).
- Junior, Salomon and Pamplona (2014) “ARIMA: An Applied Time Series Forecasting Model for the Bovespa Stock Index” *Applied Mathematics*, 5, 3383-3391 Published Online December 2014 in SciRes, <https://www.researchgate.net/publication/275214476>
- Kaastra, M. Boyd (1996) “Designing a neural network for forecasting financial and economic time series”, *Neurocomputing*, 10 (3), pp. 215-236, <https://www.sciencedirect.com/science/article/pii/0925231295000399/pdf?md5=85cce437dd80a9cc55ebaa8a43455327&pid=1-s2.0-0925231295000399-main.pdf>

- Daniel, Fabrice (2019) “Financial Time Series Data Processing for Machine Learning”, arXiv preprint arXiv:1907.03010, arxiv.org, <https://arxiv.org/pdf/1907.03010.pdf>
- Sergio, Anderson Tenório (2017) “Seleção dinâmica de combinadores de previsão de séries temporais”, Dissertação de Doutorado em Ciências de Computação - Centro de Informática, Universidade Federal de Pernambuco, Recife, <https://repositorio.ufpe.br/bitstream/123456789/25449/1/TESE%20Anderson%20Ten%C3%B3rio%20Sergio.pdf>
- Cowpertwait, P. S. P. Metcalfe, A. V. (2009) “Introductory time series with R”, Springer Science & Business Media.
- Vapnik, V. N. (1999) “An overview of statistical learning theory”, IEEE Transactions on Neural Networks, 10, 988–999, <https://ieeexplore.ieee.org/document/788640>
- PYTHON SOFTWARE FOUNDATION (2021) Python Language Site: Documentation. Página de documentação, <https://www.python.org/doc/>
- GIL, A. C. (2010) Como elaborar projetos de pesquisa, 4 ed. São Paulo, Brasil. Editora Atlas.
- Bueno, R. L. (2011) Econometria de Séries Temporais. 2 ed. São Paulo, Brasil. Editora Cengage Learning.