UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM MODELAGEM DO CONHECIMENTO

DANIEL KAZUYUKI FUGII MATSUMOTO

Estudo em Séries Temporais Financeiras utilizando Redes Neurais Recorrentes

Maceió-AL

Dezembro de 2019

DANIEL KAZUYUKI FUGII MATSUMOTO

Estudo em Séries Temporais Financeiras utilizando Redes Neurais Recorrentes

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Modelagem Computacional do Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Heitor Soares Ramos

Maceió-AL

Dezembro de 2019

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

M434e Matsumoto, Daniel Kazuyuki Fugii.

Estudo em séries temporais financeiras utilizando redes neurais recorrentes / Daniel Kazuyuki Fugii Matsumoto. – 2019.

54 f. : il.

Orientador: Heitor Soares Ramos.

Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2019.

Bibliografia: f. 53-54.

1. Análise de séries temporais. 2. Séries financeiras. 3. Domínio do tempo. 4. Redes neurais (Computação). 5. Redes neurais recorrentes. 6. *Autoregressive integrated moving average*. 7. *Long short-term memory*. 8. Mercado financeiro. I. Título.

CDU: 004.8:336.76

Folha de Aprovação

Daniel Kazuyuki Fugii Matsumoto

Estudo em séries temporais financeiras utilizando redes neurais recorrentes

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas e aprovada em 10 de dezembro de 2019.

Prof. Dr. Heitor Soares Ramos Filho
Departamento de Ciência da Computação - UFMG
Orientador

Banca Examinadora:

Prof. Dr. Osvaldo Aníbal Rosso
Instituto de Física - UFAL
Examinador externo

Prof. Dr. Leonardo Viana Pereira
Instituto de Computação - UFAL
Examinador interno

Dedico este trabalho à minha esposa Juliana, principal apoiadora das minhas decisões e meus pais que sempre estiveram dispostos a contribuir.

AGRADECIMENTOS

Agraço ao meu orientador e amigo, Prof. Dr. Heitor S. Ramos, que me deu suporte em todo o meu processo de formação acadêmica.

“It is paradoxical, yet true, to say, that the more we know, the more ignorant we become in the absolute sense, for it is only throught enlightenment that we become conscious of our limitations. Precisely one of the most gratifying results of intellectual evolution is the continuous opening up of new and greater prospects”

(Nikola Tesla)

RESUMO

Entender o comportamento das *séries temporais financeiras*, é de suma importância para conseguir prever valores futuros e tomar decisões eficientes. Esta área é frequentemente alvo de estudos e novas propostas na tentativa de sua modelagem, pois é muito volátil em resposta ao dinamismo do mercado. Essa volatividade é definida pela alta quantidade de variáveis e diferentes fontes de informações, baixa relação entre sinal e ruído, tornando falhas as suas predições ao generalizar o problema. Além de existir relações entre variáveis exóginas desconhecidas e aleatórias, capazes de forte influência na estrutura corrente dos dados. Contudo, assim como pode causar grande revés uma informação incorreta, uma informação *mais correta* pode gerar lucros proporcionais, garantindo assim que este problema seja objeto de diversos estudos durante a história. Para este objetivo, foi projetado um modelo preditivo utilizando *Redes Neurais Recorrentes* (*LSTM*) alimentada de seus valores históricos (*série temporal*) para a previsão da *tendência de variação* do valor de abertura de um ativo (*stock*). Dessa forma, uma variedade de experimentos foram executados e seus resultados analisados em relação à métricas largamente utilizadas, como os modelos autoregressivos (*ARIMA*) para definir uma base comparativa e verificar a capacidade do modelo. Os resultados obtidos foram satisfatórios, obtendo uma *acurárica* de até 74.50% ao prever o sinal do próximo valor de abertura, inferindo se uma determinada ação irá subir ou não em um ponto no futuro.

Palavras-chaves: Séries Temporais, Séries Financeiras, Domínio Temporal, Redes Neurais, Redes Neurais Recorrentes, ARIMA, *Long Short-Term Memory*.

ABSTRACT

Understanding the behavior of *financial time series* is very importance to forecast future values and make efficient decisions. This area is often the target of studies and new proposals in the attempt of its modeling, since it is very volatile in response to the dynamism of the market. This volatility is defined by the high number of variables and different sources of information, a low signal-to-noise ratio, and its predictions fail to generalize the problem. In addition to existing relations between unknown and random exogenous variables, capable of strong influence on the current structure of the data. However, just as incorrect information can cause great misfortune, *more accurate* information can generate proportional profits, thus ensuring that this problem is the subject of several studies throughout history. For this purpose a predictive model was designed using *Recurrent Neural Networks* (*LSTM*), fed from its historical dataset (*time series*) to forecast the oppening value *trend of variation* for an asset (*stock*). Thus a variety of experiments were performed and their results analyzed against the widely used metrics, like autoregressive models (*ARIMA*) to define a basis for comparisons and validate the model capacity. The results obtained were satisfactory, obtaining an *acuracy* of up to 74.50% by predicting the sign of the next opening value, inferring whether or not a given stock will rise in the future.

Keywords: Time Series, Financial Series, Time Domain, Neural Networks, Recurrent Neural Networks, ARIMA, *Long Short-Term Memory*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARIMA	Autoregressive Integrated Moving Avarage
ANN	Artificial Neural Network
MLP	Multi Layer Perceptron
DAN2	Dynamic Artificial Neural Network
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
RNN	Recurrent Neural Network
SVM	Support Vector Machine
GA	Genetic Algorithm
SA	Simulated Annealing
LSTM	Long Short Term Memory
MSE	Mean Squared Error
MAD	Mean Absolute Deviate
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive

SUMÁRIO

1 INTRODUÇÃO

Entender o comportamento das *séries temporais financeiras*, é de suma importância para conseguir prever valores futuros e tomar decisões eficientes. Esta área é frequentemente alvo de estudos e novas propostas na tentativa de sua modelagem, pois é muito volátil em resposta ao dinamismo do mercado.

Essa volatividade é definida pela alta quantidade de variáveis e diferentes fontes de informações, baixa relação entre sinal e ruído, tornando falhas as suas predições ao generalizar o problema. Além de existir relações entre variáveis exóginas desconhecidas e aleatórias, capazes de forte influência na estrutura corrente dos dados.

Contudo, assim como pode causar grande revés uma informação incorreta, uma informação *mais correta* pode gerar lucros proporcionais, garantindo assim que este problema seja objeto de diversos estudos durante a história.

Dessa forma, este trabalho almeja utilizar os métodos existentes, relacionados ao *domínio do tempo*, por conta de seus atributos estarem sequencialmente divididos em intervalos regulares temporais. Posteriormente, os dados utilizados na análise anterior serão utilizados para tentar prever a variação de valores, onde é considerado se o próximo valor de abertura de um determinado ativo terá uma variação positiva ou não em relação ao valor corrente e não o valor absoluto como é o foco da maioria dos estudos nesta área.

Para esta finalidade, este trabalho fará uso de modelos consolidados como eficientes pela literatura, como o modelo **ARIMA** e redes **LSTMs** em diversas configurações sobre o mesmo conjunto de dados, pertencentes a uma *série temporal financeira* sobre um ativo de mercado. Sequentemente será comparada a *eficácia* dos métodos utilizados através da *variação* prevista e os valores reais.

No capítulo seguinte (Capítulo 3) serão apresentados alguns trabalhos relacionados, suas similaridades e distinções.

Posteriormente serão apresentadas as metodologias utilizadas (Capítulo 4), seus resultados (Capítulo 5) e será discorrido no capítulo final (Capítulo 6).

1.1 Motivações

Ao longo dos últimos anos o estudo de séries temporais financeiras utilizando redes recorrentes tem se demonstrado efetiva, dentre os diversos estudos com LSTMs, muitos demonstram que uma previsão assertiva dos dados é uma tarefa complexa de se adquirir. Contudo, apesar de seus diversos estudos envolvendo essas redes para determinar um valor concreto, este trabalho tem a motivação de usar estas redes para avaliar a variação de uma determinada ação de mercado e predizer o aumento ou redução do próximo valor de abertura em relação ao corrente.

Para essa finalidade foi utilizado várias observações realizadas com diferentes configurações de LSTMs, como pode ser observado na Figura 1, que ao longo do processo de aprendizagem e predição, demonstraram relevante característica em modelar a série temporal, gerando uma boa representação da tendência de mercado. Através destas observações, construiu-se a hipótese de que uma LSTM, devidamente treinada, é capaz de fazer previsões mais assertivas sobre a tendência de variação de abertura de ações da bolsa de valores, ao contrário de tentar predizer o próximo valor real de abertura.

1.2 Objetivos

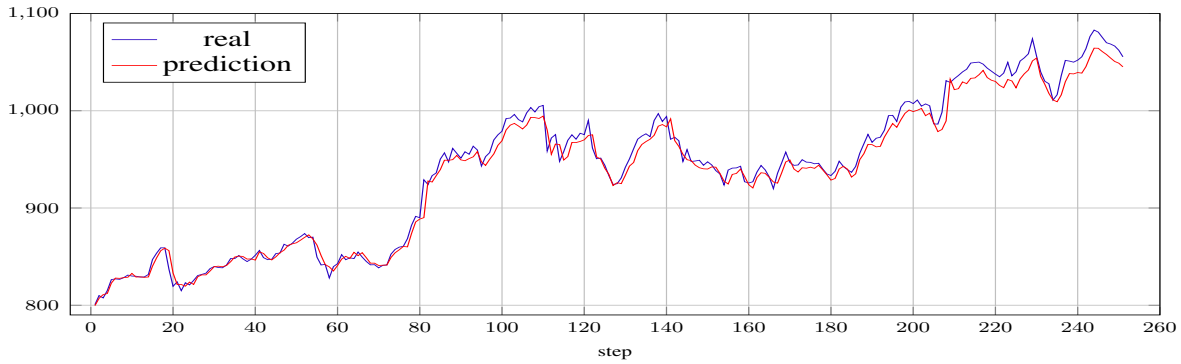
Desenvolver uma metodologia preditiva da variação do valor de abertura de um determinado ativo através do uso de uma *LSTM* alimentada de seus valores históricos (*série temporal*). Onde a variação é baseada no sinal sobre o valor corrente e o predito. Para isso, este trabalho almeja definir parâmetros bases de avaliação de desempenho através de técnicas e modelos existentes (**LSTMs** e **ARIMA**), comparando-os através das previsões realizadas acerca da tendência de variação do valor de abertura. Para que esse processo seja possível, serão construídas diferentes configurações dos modelos supracitados e serão utilizadas como parâmetros de treinamento as séries temporais financeiras de diferentes ações, dessa forma será possível fazer comparações e testes entre as diferentes propriedades.

1.3 Justificativa

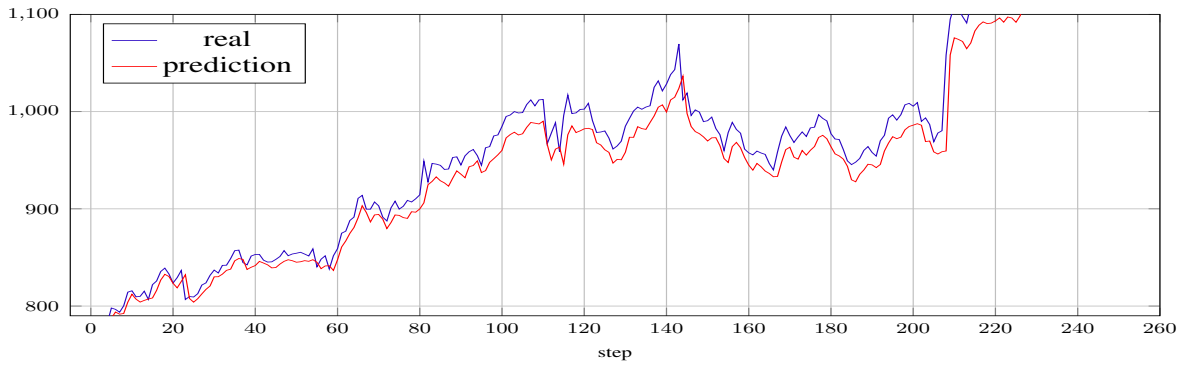
Apesar de existirem diversos estudos utilizando diferentes metodologias de predição, dentre elas as redes neurais, estes em sua grande maioria tem o objetivo de prever valores futuros. Este trabalho difere neste objetivo, onde visa prever a tendência de variação do valor de abertura utilizando dados históricos, conjunto de atributos da série temporal financeira, demonstrando a capacidade de modelar em certo grau de acurácia o comportamento da tendência de variação do ativo através de uma *LSTM*.

Como podemos observar na Figura 1, as diferentes configurações das LSTMs conseguem representar com certo grau de acurácia a tendência de variação do valor de abertura de um ativo de mercado, mesmo divergindo do valor exato em valores absolutos.

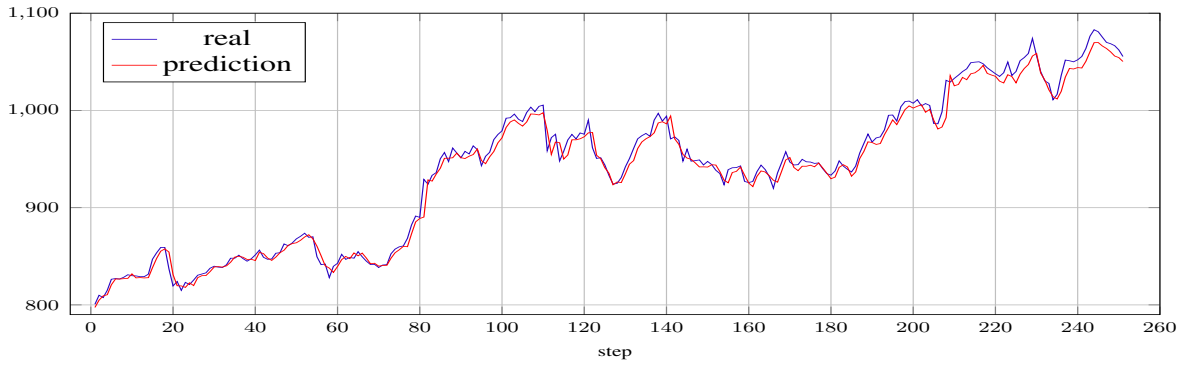
Dada essa característica, surgiu a hipótese da possibilidade de utilizar estes valores previstos como índice comparativo do valor corrente de abertura de um ativo para predizer o sinal de sua variação no próximo período. Analogamente aos estudos tradicionais em predizer o valor absoluto, este método demonstra ter forte relação dos dados históricos com a tendência de variação futura.



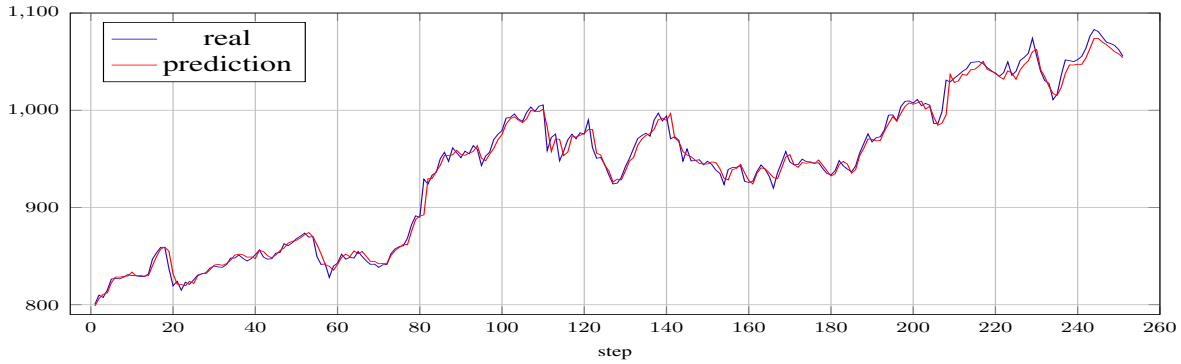
(a) Exemplo de previsão lstm-(20,1)



(b) Exemplo de previsão lstm-(30,1)



(c) Exemplo de previsão lstm-(40,1)



(d) Exemplo de previsão lstm-(50,1)

Figura 1: Diferentes configurações de LSTMs com janelas de 1 dia

2 CONCEITOS PRELIMINARES

Neste capítulo serão apresentadas conceitos gerais necessários para uma sólida fundamentação metodológica utilizado neste trabalho, assim como quais técnicas escolhidas e suas motivações. Será subdividido em domínios dos dados, onde estes são relacionados pelo tempo (Sessão 2.1), frequência (Sessão 2.2) ou através de representações simbólicas, consecutivamente serão apresentadas as principais ferramentas utilizadas (Sessão 2.3) e objeto das análises (Sessão 2.4).

2.1 Domínio do Tempo

Domínio do Tempo (***Time Domain***) é um termo utilizado para analisar funções matemáticas, sinais, séries temporais em relação ao tempo. Possui valores conhecidos em todos os intervalos observados.

2.1.1 Séries Temporais

Séries temporais (***Time Series***) são sequências de observações ao longo de intervalos, geralmente uniformes, com principal característica na dependência temporal (TIME. . . , 2018). Genericamente, podem ser decompostas em:

- ***Tendência (Trend):*** é a representação de variações de baixa frequência em uma série temporal, com suas médias e altas flutuações devidamente filtradas;
- ***Sazonalidade (Seasonality):*** é a identificação de padrões regulares de repetição de altos e baixos relacionados a fatores sazonais;
- ***Ciclo (Cycle):*** estruturas de repetição, mais ou menos regulares, relacionadas a uma sequência de tendências.

Muitas vezes são representadas por componentes matemáticos determinísticos e estocásticos, podendo ser estacionária ou não, dependendo da especificidade das observações. As principais áreas de estudo abordam entender analiticamente a estrutura que gerou a série ou efetuar uma previsão a partir de um modelo matemático gerado da mesma, este último escolhido para este trabalho.

- **Exemplos:**
  - Valores mensais de CO_2 [Figura 2];
  - Quantidade de assaltos armados em uma cidade durante um período [Figura 3];
  - Índices diários de ações na bolsa de valores [Figura 4].

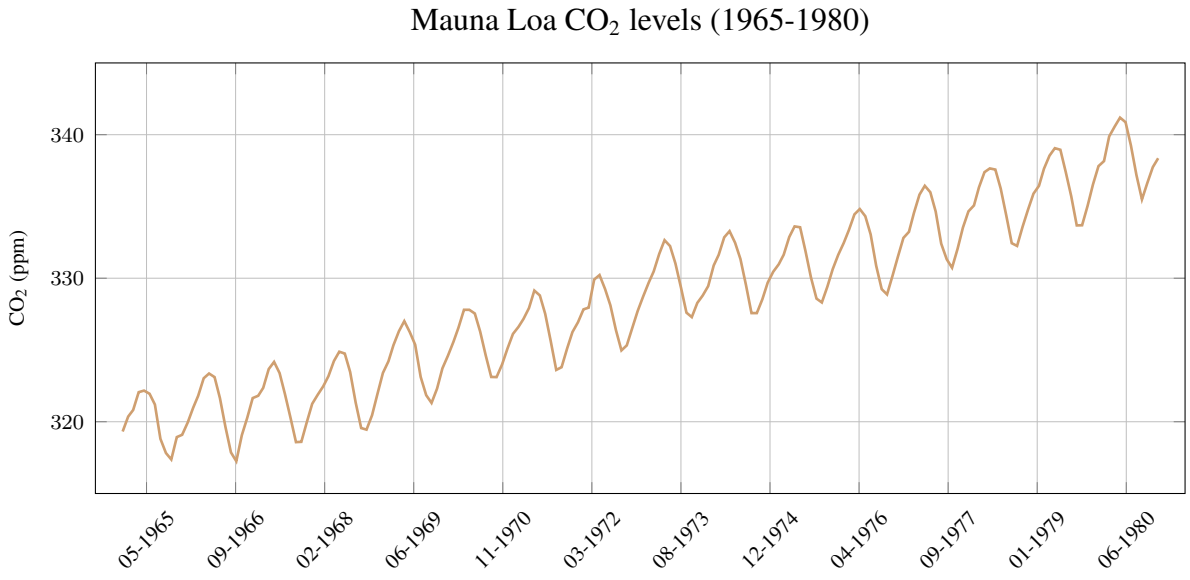


Figura 2: Exemplo de Série Temporal: Níveis de CO_2

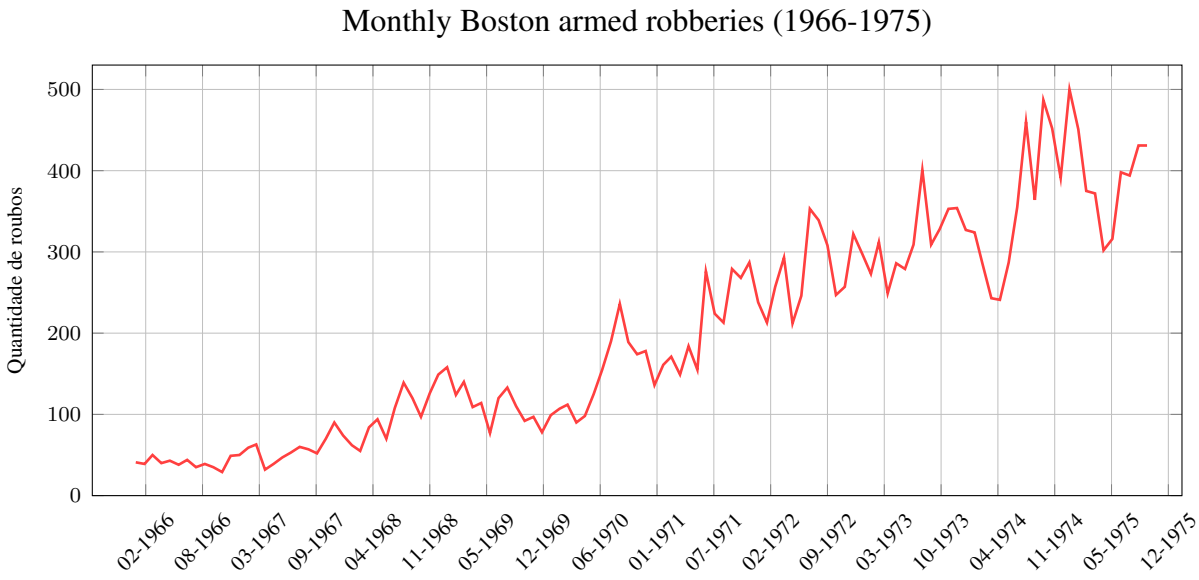


Figura 3: Exemplo de Série Temporal: Quantidade de Roubos

2.1.1.1 ARIMA

O modelo *Auto-Regressivo Integrado de Médias Móveis* (**ARIMA**) é muito utlizado na compreenção de séries temporais ou predição de um ponto no futuro (***forecast***). Podem ser aplicados a casos onde não há evidencias de não estacionariedade nos dados, utilizando a diferenciação, correspondente à parte *integrada* do modelo, para eliminar-la.

Basicamente pode ser descrita através de suas partes:

- **AR (Autoregressive):** essa parte auto-regressiva do modelo indica que a variável de interesse sofre uma regressão em seus prévios valores;

Figura 4: Exemplo de Série Temporal: Mercado Financeiro

- **I (Integrated):** parte responsável em indicar que os valores de dados foram substituídos com a diferença entre seus valores correntes e os anteriores, podendo acontecer mais de uma vez;

- **MA (Moving Average):** parte responsável em indicar que o *erro de regressão* é uma *combinação linear* dos termos dos erros, cujos valores ocorrem simultaneamente em diferentes momentos passados.

Seus modelos não sazonais são geralmente denotados como **ARIMA(p, d, q)**, onde p, d e q são números inteiros não negativos. E nos modelos sazonais, são habitualmente denotados por **ARIMA(p, d, q)(P, D, Q)**$_m$.

- **Definição:**

  - ***p*:** é a ordem (passos anteriores) do modelo, mesmo propriedade observada numa janela deslizante de dados, utilizada na Sessão 4.1 de pre-processamento dos dados;
  - ***d*:** é o grau de diferenciação, ou seja, número de vezes em que os dados passados foram subtraídos;
  - ***q*:** ordem do modelo para as médias móveis;
  - ***P*, *D* e *Q*:** termos da parte sazonal do modelo;
  - m**:** número de períodos em cada temporada (*season*).

Tabela 1: Exemplos ARIMA(p, d, q)

Configuração	Descrição
ARIMA(0, 1, 0)	$X_t = X_{t-1} + \varepsilon_t$, ou passeio aleatório simples
ARIMA(0, 1, 0)	$X_t = c + X_{t-1} + \varepsilon_t$, ou passeio aleatório com desvio
ARIMA(0, 0, 0)	modelo de *ruído branco*
ARIMA(0, 1, 2)	modelo de *Holt* amortecido
ARIMA(0, 1, 1)	modelo básico de suavização exponencial
ARIMA(0, 2, 2)	$X_t = 2X_{t-1} - X_{t-2} + (\alpha + \beta - 2)\varepsilon_{t-1} + (1 - \alpha)\varepsilon_{t-2} + \varepsilon_t$

Fonte: Casos especiais do modelo ARIMA, 2018.

É válido ressaltar que o modelo ARIMA(0, 2, 2) descrito na Tabela 1 é equivalente ao método linear de *Holt* com erros aditivos ou suavização exponencial dupla.

2.2 Domínio de Frequência

O Domínio da Frequência (***Frequency Domain***) analisa funções matemáticas ou sinais com respeito à frequência. Em contraste, uma representação do domínio do tempo apresenta a variação do sinal sobre o tempo, enquanto o domínio da frequência quantifica quanto do sinal reside em cada faixa de frequência.

Neste domínio pode conter informações sobre deslocamentos de fase (*phase shift*), usado para recombinar os componentes da frequência e recuperar o sinal temporal original.

2.3 Redes Neurais Artificiais

As Redes Neurais Artificiais (***Artificial Neural Network***) são modelos matemáticos inspirados pelo processo funcional do sistema nervoso central biológico, possuindo a capacidade de aprendizado de máquina, assim como reconhecimento de padrões. Geralmente são representados por *coleções de nerônios artificiais interconectados* [Figura 5], simulando o comportamento real esperado.

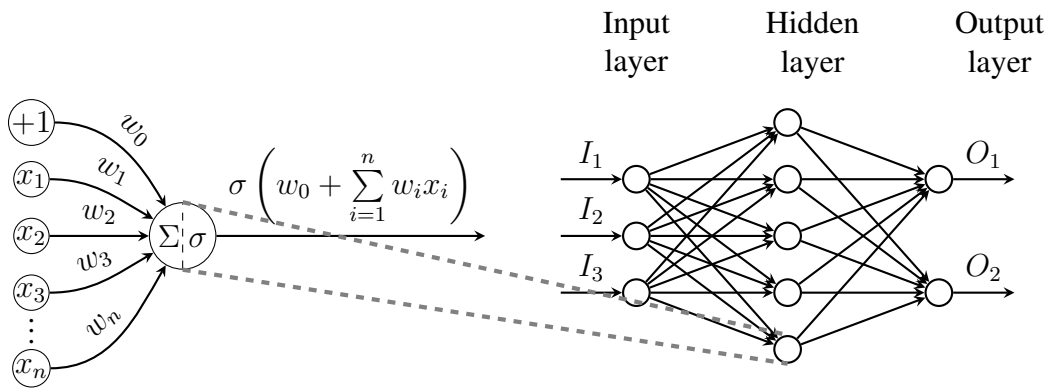


Figura 5: Exemplo de Rede Neural Artificial

Essa concepção foi idealizada como uma tentativa de representação matemática análoga ao processo biológico de um cérebro, onde seus componentes básicos de quantificação da informação são equivalentes aos *neurônios* e suas reações químicas representadas através de *funções de ativação*.

Cada um dos *neurônios*, que são as unidades de processamento elementares de uma rede neural artificial, tem a função de simular um comportamento biológico através de um *peso*, responsável por contribuir na modelagem de um atributo avaliado dentre as entradas de dados $(x_1, x_2, ..., x_n)$ e posteriormente usar uma função de transferência para transportar o dado para as próximas unidades elementares conectadas diretamente. A definição desses *pesos* acontece na etapa de treinamento, onde são feitos cálculos sucessivos, durante suas iterações, na tentativa de minimizar o erro através de um resultado esperado (*treino supervisionado*).

- **As RNAs podem ser decompostas genericamente em camadas:**
  - *Entrada (input layer):* responsável pela entrada de dados no modelo $(I_1, I_2, ..., I_n)$, variando a quantidade de atributos processados no problema com o mesmo número de ***neurônios*** de entrada;
  - *Processamento (hidden layer):* responsável por gerar a representação autonômica da informação, variando em quantidades de ***neurônios*** por camada e/ou quantidade de camadas;
  - *Saída (output layer):* responsável pela apresentação dos dados de forma estruturada, variando a quantidade de ***neurônios*** baseado no formato do resultado.

Estes modelos tem grande aplicação em tarefas complexas de se resolver com programação tradicional, pela capacidade de modelar problemas não lineares, incluindo *visão computacional*, *aprendizado e tomada de decisões*, *reconhecimento de escrita*, *diagnósticos médicos*, etc.

Com diferentes funcionalidades e aplicações, essas redes se diferenciam pela forma de construção de seu modelo, variando a forma de aprendizado e representação autonômica da informação, além da apresentação de resultados ou estimativas.

Tradicionalmente, são compostas por um fluxo de dados unidirecional, não existindo ciclos ou repetições entre seus neurônios, denominadas redes *Feed-Forward*. Geralmente definidas como uma função de sua entrada, pois não contém estados internos além de seus pesos.

- **Exemplos:**
  - *Perceptron Multicamadas [Figura 5]*;
  - *Auto-encoderes*;

- *Redes Neurais Convolucionais*;
- *Redes Neurais Recorrentes [Figuras 6, 7, 8].*

2.3.1 Redes Neurais Recorrentes

As redes recorrentes (***Recurrent Neural Networks***) são classes de redes neurais especialmente úteis para processar dados sequenciais. Elas se diferenciam das redes tradicionais por conterem um *loop de feedback*, onde o resultado do passo anterior é realimentada na rede para afetar o passo subsequente, sucessivamente, como podemos observar na *Figura 6*.

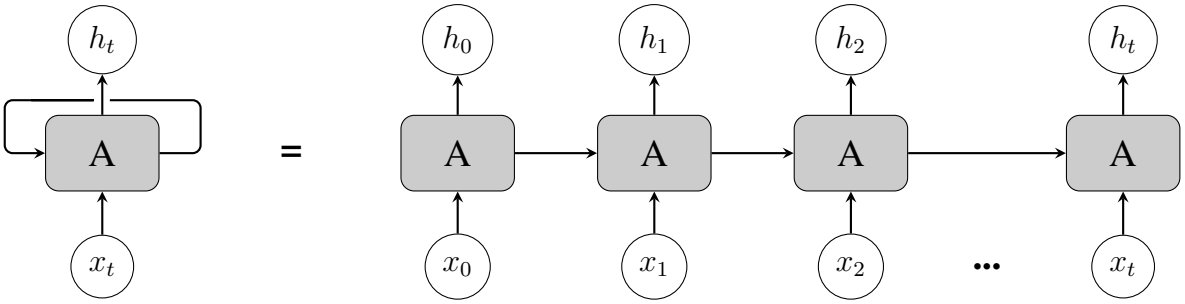


Figura 6: Rede Neural Recorrente

Em contraste com as redes tradicionais, que produzem um modelo estático dos dados capaz de aceitar novos exemplos para classificar/agrupar com certa precisão, as redes recorrentes geram modelos dinâmicos capazes de captar as relações sequencias dos exemplos anteriores e integrar essas informações no processo corrente.

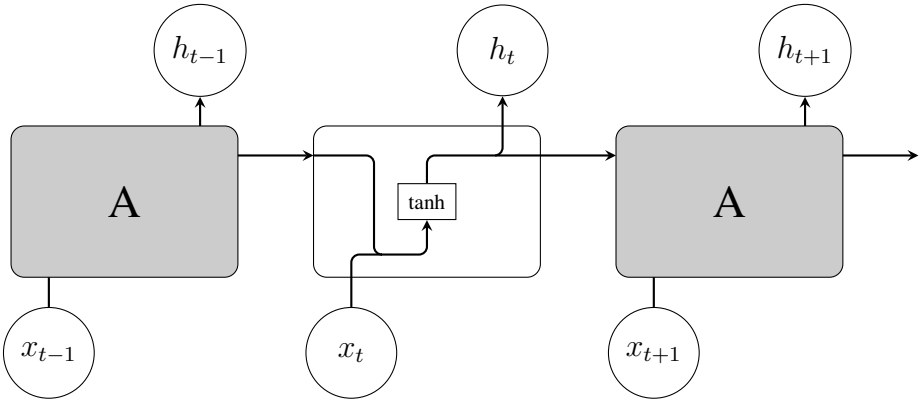


Figura 7: Rede Neural Recorrente detalhada

Suponha x como um conjunto de valores provenientes de medições realizadas para avaliar os níveis de CO_2 segmentados em meses durante o período de 1965 até 1980. Para observarmos o valor coletado do mês t, obtemos x_t. Utilizando a rede recorrente da *Figura 7*, podemos observar que o resultado no momento t necessita de processos realizados no momento $t-1$, posteriormente seu processamento será usado no momento subsequente $t+1$. Podendo ser simplificado como uma simples multiplicação de x_t e h_{t-1} passadas por uma função de ativação $tanh$ (Figura 7).

Em outras palavras, a rede neural adquire a capacidade de *memória de curto prazo* com a persistência de informações no processamento de dados sequenciais. Por essa característica, são muito utilizadas no estudo de dados sequenciais e listas.

- ***Alguns exemplos de estudos realizados com redes recorrentes:***
  - *Reconhecimento de voz;*
  - *Tradução;*
  - *Legenda de imagens.*

Apesar do conceito de reciclar informações anteriores para serem usadas no processo corrente ser muito atrativo, essas redes tendem a apresentar problemas na magnitude dos gradientes, sumindo ou explodindo, em dependências de longo termo, ocasionando falha na modelagem dos dados.

- ***Alguns exemplos de variações das redes recorrentes:***
  - *Redes de Hopfield;*
  - *Máquinas de Boltzmann;*
  - *Redes Long Short Term Memory.*

2.3.1.1 Redes LSTM (Long Short Term Memory)

As *LSTMs* são variações especiais de redes recorrentes, onde são capazes de aprender dependências de longo termo, relembrando informações por períodos mais longos e não apresentando assim a problemática dos gradientes.

Assim como as redes recorrentes convencionais, as *LSTMs* também tem sua estrutura atrelada a cadeias de repetições, contudo divergem em complexidade, observado nas *Figuras 7 e 8*. A principal diferença é a presença de portões (***GATES***) com a capacidade de remover ou adicionar informações ao seu estado, em outras palavras são uma forma opcional de permitir informações e são compostas de uma camada de rede sigmóide (σ).

Sua capacidade de manter informações fora do fluxo convencional da rede garante a habilidade de modular as dependências de longo termo através do processo iterativo de aprendizagem. Onde inicialmente é decidido quais informações vão ser *"esquecidas"* através do portão sigmóide e para cada h_{t-1} e x_t é calculado um número entre 1 e 0 representando a *"taxa de informação esquecida"*. Posteriormente é decidido quais informações vão ser armazenadas em seu novo estado, passando por um portão sigmóide para definição de quais valores atualizar e em seguida uma camada $tanh$ cria um vetor de novos valores candidatos que podem ser adicionados ao estado. Após as etapas anteriores, seus resultados são combinados para criar uma atualização

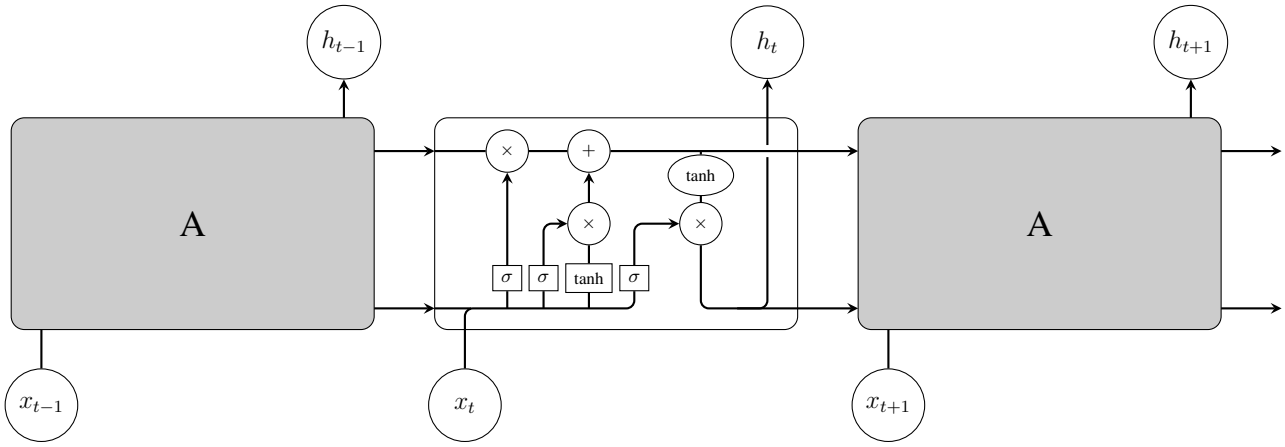


Figura 8: LSTM Unit

no estado interno da unidade *LSTM*. O último procedimento é executar um portão sigmóide para decidir quais partes do estado vão ser emitidas e em seguida passar o estado da unidade através de uma $tanh$ (normalização de valores entre -1 e 1) obtendo o resultado do estado a ser emitido.

- ***Alguns exemplos de estudos realizados com LSTMs:***
  - *Processamento de linguagens naturais;*
  - *Geração automática de texto;*
  - *Séries temporais*

2.4 Mercado Financeiro

O mercado financeiro é onde a negociação e transação de ativos, mercadorias e câmbio são realizadas. São considerados parte inerente e vital de qualquer economia, pois sua movimentação pode ser traduzida como oportunidades econômicas, geralmente responsável pelas variações de mercado.

O termo *mercado financeiro* é utilizado por causa do processo de compra e venda, onde investidores negociam tais bens entre si, seguindo o principio de obtenção de lucro através deste processo iterativo. Contudo, este tarefa se torna laboriosa uma vez que a variação de valor negociado é extremamente sensível e não é apenas influenciado pela demanda, mas é de conhecimento que variáveis exógenas tem peso considerável nestas variações, como política territorial, notícias, boatos e eventos naturais.

Dentre as diversas categorias negociadas no mercado financeiro, as *ações* representam o direto à propriedade e lucros de uma determinada empresa. São emitidos no momento em que a empresa queira abrir seu capital ao público, para este ser negociado na *bolsa de valores*. Além da relação de demanda pela ação, seu valor também é influenciado pelo reflexo da capacidade financeira/econômica da empresa e, assim como supracitado, interferências exógenas diversas.

Ao relacionar as variações de uma determinada *ação* com o ***domínio do tempo***, obtemos uma ***série temporal*** dos dados conhecida como *série financeira* dos dados históricos, utilizadas como objeto de estudos neste trabalho para modelar o comportamento dos dados e predizer valores futuros de flutuação da *ação*.

Em relação à esse objetivo, baseado na complexidade e dinamismo inato do problema, um constante debate sobre a possibilidade de prever mudanças nos valores das ações provocou vertentes distintas em teorias e hipóteses em como abordar a modelagem destas variações. Algumas delas aceitam a hipótese do **Mercado Eficiente** atribuindo que o preço atual de um ativo sempre reflete a toda informação previamente existente instantaneamente naquele momento, definindo que não é possível fazer uso de quaisquer informações previamente conhecidas para prever as variações futuras.

Analogamente, existe a hipótese do **Passeio Aleatório**, inferindo que a variação de preço de uma ação independe de seu histórico, mas sim, somente atributos e eventos futuros são responsáveis diretamente pela mudança de seu valor.

Contudo, uma outra vertente de pesquisadores alegam por meio de experimentos, que essas variações podem ser previstas até certo nível. Dessa forma, foi desenvolvido uma variedade de métodos diferentes para modelagem e predição dessas séries financeiras, sendo muito utilizadas ao longo das últimas décadas.

3 TRABALHOS RELACIONADOS

Atualmente, existem diversos estudos sobre *séries temporais* utilizando uma grande variedade de métodos de análise, como o *ARIMA* e *LSTMs*. Em sua grande maioria, os estudos presentes nesta área se concentram em relacionar os dados históricos com um valor futuro, objeto almejado dentro do mercado financeiro por ser uma informação extremamente relevante a tomada de decisões. Porém, existem duas vertentes na abordagem sobre a possibilidade de modelar o movimento das ações.

A primeira parte do pressuposto de que não existe relação entre os dados históricos e os dados futuros, consequentemente definindo a hipótese de que não é possível prever mudanças dos valores através de informações previamente conhecidas, nesta vertente existem dois importantes estudos que contribuem para este objetivo, sendo elas a *hipótese do Mercado Eficiente* (FAMA; MALKIEL, 1970) e a *hipótese do Passeio Aleatório* (MALKIEL, 1973). Apesar de que estes estudos estabeleçam a impossibilidade de predição do valor de uma ação, existem trabalhos que validam a estratégia de investimento utilizando uma metodologia de tomadas de decisões aleatórias através de comparações com métodos tradicionais estatísticos.

A segunda vertente defende através de experimentos da possibilidade de predição de valores à partir de seus dados históricos até certo nível, desta forma este trabalho utiliza deste pressuposto da existência da relação destas informações históricas com os movimentos futuros da tendência de variação de um determinado ativo.

Nesta área de pesquisa, o uso de inteligência computacional é bastante difundido pela sua grande capacidade de processamento de objetos complexos utilizando os mais diversos modelos algorítmicos, como *Algoritmos Genéticos* (QIU; SONG, 2016), que são implementados como uma simulação em que uma população de representações abstratas da solução é selecionada através de processos iterativos, chamados de *gerações*, que validam a adaptação sobre cada solução corrente da etapa, recombinando ou mutando os indivíduos para a próxima *geração*, contudo o principal problema desta metodologia é que se faz uso de transições probabilísticas e não regras determinísticas.

Outro método muito utilizado é a *Máquina de Vetor de Suporte* (**SVM**) (KIM, 2003), onde estes modelos se diferenciam deste trabalho pela característica de classificação e seleção. A *SVM*, através de processos lineares e não probabilísticos, define entre os resultados uma separação conhecida como *hiperplano* entre os dados de duas classes, essa distância entre o *hiperplano* e o primeiro ponto de cada classe costuma ser chamada de margem. A SVM coloca em primeiro lugar a classificação das classes, definindo assim cada ponto pertencente a cada uma delas e em seguida maximiza a margem.

Uma metodologia frequentemente utilizada são as *Redes Neurais*, difundidas pela sua capacidade analítica e diversidade em sua construção, definindo aspectos únicos dependendo

do propósito. Por estes fatores, muitos trabalhos focam na comparação entre diferentes técnicas existentes, como a comparação entre *ARIMA x LSTM* (SIAMI-NAMINI; NAMIN, 2018), que foi utilizada como base referencial e motivacional para desenvolver os estudos presentes neste trabalho, contudo divergindo na metodologia de avaliar a *tendência de variação* de um ativo, como observado em (FARIA et al., 2009), ao contrário de um valor absoluto.

Alguns estudos se baseiam na construção de diferentes composições das *Redes Neurais* (HEDAYATI; HEDAYATI; ESFANDYARI, 2016). Nestes trabalhos podemos observar a tentativa de prever resultados com maior nível de *acurácia* na tentativa de ultrapassar os métodos lineares e não lineares tradicionais. Podemos observar um estudo dirigido a este tema (GURESEN; KAYAKUTLU; DAIM, 2011), onde o autor utiliza a comparação de *MSE* e *MAD* entre **MLPs**, **DAN2** e **GARCH** utilizando valores reais da ***NASDAQ Stock Exchange***.

Dentro do campo de estudos das *Redes Neurais*, muitos trabalhos utilizando redes profundas (**Deep Neural Networks**) obtiveram resultados satisfatórios na análise de dados complexos, como textos e notícias em linguagem natural, colaborando com a previsão de valores futuros nos ativos. Podemos citar (...).

Nas sessões abaixo (Sessões 3.1 e 3.2), serão apresentados uma breve observação sobre os métodos mais comuns encontrados na literatura nos diversos possíveis usos, mantendo o foco de apresentar de maneira geral trabalhos relacionados ao propósito deste trabalho.

3.1 ARIMA

Este método é frequentemente utilizado para predizer valores futuros em *séries temporais*, variando em predizer volume de aplicações para empregos (XIAOGUO; YUEJING, 2009), simulações de atenuação de chuva (YANG et al., 2013), previsões da velocidade do vento (ZHE; YU; ZIJUN, 2014) ou índice de clareza (HASSAN, 2014). Em teoria, esse modelo pertence a classe mais geral de previsões, trabalhando com *dados estacionários* através da *diferenciação* se houver necessidade.

Neste trabalho, este modelo será usado como base comparativa para os modelos subsequentes (*LSTMs*) utilizando o valor do *erro quadrático médio* (**MSE**) das previsões realizadas.

3.2 LSTM

Nos últimos anos, diferentes estudos envolvendo *redes neurais recorrentes* estão sendo utilizados para modelar dados temporais. Podemos citar a previsão de valores do *Bitcoin* utilizando *LSTM* (MCNALLY; ROCHE; CATON, 2018) ou previsão de *eventos emergenciais* (CORTEZ et al., 2018) como uma evolução natural da Sessão 3.1 para melhorar a acurácia de previsão.

Muitos estudos, como estudos comparativos de previsões de surtos da *influenza* (ZHANG;

NAWATA, 2017), fazem comparação dos diversos métodos e apresentam a LSTM como alternativa viável.

3.3 Séries Financeiras

Atualmente, os estudos que fazem uso da inteligência computacional para predição de valores de ativos (*stocks*) no mercado financeiro tem progredido através de trabalhos utilizando os mais variados métodos.

Em contrapartida, temos trabalhos que fazem uso de *redes neurais Bayesianas* para tentar prever os valores de fechamento (*CLOSE*) de ativos utilizando os dados históricos e indicadores técnicos como parâmetros de entrada (TICKNOR, 2013), visando minimizar o *overfitting* e o *overtraining* para melhorar a qualidade das previsões.

Apesar dos diversos estudos fazendo uso de *ANNs* na tentativa de predição de valores futuros, muitos deles fazem uso de soluções híbridas para melhorar a qualidade das previsões (QIU; SONG; AKAGI, 2016), como o uso de *algoritmos genéticos* e **SA** (*Simulated Annealing*).

No trabalho de (FARIA et al., 2009) é utilizado uma *ANN* com *backpropagation* para predizer o valor do sinal de um ativo, utilizando diversas configurações. Esta metodologia compara seus resultados contra um método estatístico para predições que utiliza as variações sazonais como um coeficiente (***adaptive exponential smoothing method***), sendo que este é permitido a flutuação ao decorrer do tempo para refletir mudanças significativas.

Neste trabalho, é proposto um método baseado em redes *LSTMs* para prever a *tendência de variação* (movimento de preços) do valor de abertura de um ativo, fazendo uso somente de seus dados históricos (*série temporal*) e posteriormente comparando seus resultados com metodologias consolidadas (*ARIMA*).

4 METODOLOGIA

Neste estudo foram utilizados dados do mercado financeiro para ações (*stocks*) das empresas ***Amazon***, ***Ebay***, ***Facebook***, ***HP*** e ***Google*** (**Alphabet Inc. A**), entre o período de Janeiro de 2010 até Dezembro de 2017, disponibilizados pela empresa ***Quantiacs***, ao qual possui dados do mercado financeiro durante os últimos 25 anos de algumas empresas/mercados.

Os dados utilizados foram pré-processados em ativos de informações refinadas e descritas com mais detalhes na sessão subsequente (Sessão 4.1), após esta etapa foi necessário criar métricas de referência para comparação com as pesquisas realizadas neste estudo. Para esta finalidade foram desenvolvidas ***Redes Neurais Recorrentes*** (**RNNs**), mais especificamente ***Long Short Term Memory*** (**LSTM**) com diferentes quantidades de unidades e janelas de dados.

Em contrapartida, foi utilizado o modelo ***ARIMA*** (**Auto Regressive Integrated Moving Average**) com diferentes configurações para formalizar uma base comparativa entre as técnicas utilizadas através da diferença entre os erros quadráticos médios.

4.1 Preparação dos Conjuntos de Dados

Foram escolhidos as ações das empresas ***Amazon***, ***Ebay***, ***Facebook***, ***HP*** e ***Google*** (**Alphabet Inc. A**), segmentadas entre as datas de Janeiro de 2010, início do período de transações financeiras deste ano, até Dezembro de 2017, último dia de transações neste período (Tabela 2). Neste conjunto de dados temporais de cada ação, serão utilizadas as seguintes informações: valor de abertura, maior valor atingido, menor valor atingido, valor de fechamento e volume de ações (OPEN, HIGH, LOW, CLOSE, VOL respectivamente). Posteriormente foi eliminada a informação referente à data, uma vez que não se faz necessário o uso desse atributo utilizando janelas de visualização (**sliding window**).

Tabela 2: Google stock (2010-2017)

DATE	OPEN	HIGH	LOW	CLOSE	VOL
2010-01-04	313.823	315.070	312.432	313.688	3912010.0
2010-01-05	314.234	314.234	311.081	312.307	6009698.0
2010-01-06	312.998	313.243	303.483	304.434	7953295.0
2010-01-07	305.005	305.305	296.621	297.342	12823204.0
2010-01-08	296.396	301.926	294.849	301.311	9439427.0
⋮	⋮	⋮	⋮	⋮	⋮
2017-12-28	1062.250	1064.840	1053.380	1055.950	966131.0
2017-12-29	1055.490	1058.050	1052.700	1053.400	1074892.0

Fonte: Quantiacs (Alphabet Inc. A).

As janelas de visualizações são a base necessária para transformar a série temporal dos dados em problemas de aprendizado supervisionados por regressão ou classificação, desde que a

ordem das linhas seja preservada (ordem cronológica). A quantidade de linhas define o ***time step*** e neste estudo foram utilizados *[1, 5, 10, 20, 24, 48, 72]* para definir janelas de: um dia, uma semana, duas semanas, quatro semanas, um mês, dois meses e quatro meses respectivamente, baseado em dias úteis do mercado financeiro.

Nesta etapa, é necessário levar em consideração que o tamanho das janelas influencia diretamente o tamanho dos dados de treinamento assim como a data inicial de treino. A Tabela 3 é um exemplo de janela de cinco dias.

Tabela 3: Exemplo de Janela de visualização

OPEN_{t-5}	HIGH_{t-5}	LOW_{t-5}	CLOSE_{t-5}	VOL_{t-5}	...	CLOSE_{t-1}	VOL_{t-1}
313.823	315.070	312.432	313.688	3912010.0	...	301.311	9439427.0
314.234	314.234	311.081	312.307	6009698.0	...	300.855	4419174.0
312.998	313.243	303.483	304.434	7953295.0	...	295.535	9696792.0
305.005	305.305	296.621	297.342	12823204.0	...	293.838	12985150.0
296.396	301.926	294.849	301.311	9439427.0	...	295.220	8474657.0
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

Dessa forma, cada janela possui uma quantidade diferente de atributos, assim como mostrado na tabela abaixo (Tabela 4).

Tabela 4: Quantidade de atributos por janela de visualização

Janela	Quantidade de atributos
1	5
5	25
10	50
20	100
24	120
48	240
72	360

Consequentemente os dados foram normalizados para que as diferentes escalas dos atributos permaneçam em uma escala proporcionalmente comum, definidas entre *[0, 1]*, apresentados na Tabela 5.

Tabela 5: Exemplo de janela normalizada

OPEN_{t-5}	HIGH_{t-5}	LOW_{t-5}	CLOSE_{t-5}	VOL_{t-5}	...	CLOSE_{t-1}	VOL_{t-1}
0.109	0.108	0.111	0.110	0.121	...	0.095	0.310
0.109	0.107	0.109	0.108	0.193	...	0.095	0.480
0.108	0.106	0.101	0.099	0.259	...	0.089	0.319
0.099	0.097	0.093	0.091	0.426	...	0.087	0.431
0.089	0.093	0.090	0.095	0.310	...	0.088	0.277
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

4.2 LSTMs

Em relação as Redes Neurais Recorrentes, foi escolhida a LSTM pela característica de preservar o gradiente de treinamento, muito adequadas e estudadas na literatura para classificar, processar e fazer previsões em séries temporais. Para todos as redes LSTMs utilizadas neste trabalho foi adotado o padrão: camada de entrada, uma camada de processamento, uma camada opcional (***dropout***) e a camada de saída. Em cada unidade LSTM da camada de processamento é utilizado a função de ativação ***tanh*** e a função de ativação recorrente sigmóide.

Dessa forma foram adotadas diferentes quantidades de unidades presentes nas diferentes redes, como forma de estudo mais abrangente, possuindo *[20, 30, 40, 50]* unidades respectivamente na camada de processamento.

Para cada rede concebida dessa forma, serão utilizadas as diferentes janelas de dados geradas na sessão anterior (Sessão 4.1), totalizando 28 redes distintas. Além dessas redes supracitadas, serão desenvolvidas mais 28 redes semelhantes, diferindo na presença de uma camada de ***Drop Out*** (5%) para testar o ***overffiting***. Contudo será utilizado a mesma quantidade de épocas e tamanho dos lotes de treinamento (100 *epochs* e *batch size* de 100 elementos) em todos os casos.

No treinamento de cada rede foi utilizado o período de Janeiro de 2010 até Dezembro de 2016 e para validação o restante do período do ano de 2017. Cada janela possui uma quantidade de dados disponíveis para treinamento diferente, apresentados na tabela abaixo (Tabela 6)

Tabela 6: Quantidade de treinamento e validação das diferentes janelas

Janela	Qtd. treinamento	Qtd. validação
1	1761	251
5	1757	251
10	1752	251
20	1742	251
24	1738	251
48	1714	251
72	1690	251

Sequentemente a etapa de validação acontece levantando dados estatísticos de previsão e gerando uma matriz confusão para obtenção da métrica base deste modelo, além do **MSE** para comparaçao com o modelo *ARIMA*. Assim como os resultados do treinamento, os resultados da validação podem ser observados na sessão de Resultados.

4.3 ARIMA

Os testes com este modelo foram executados utilizando as seguintes combinações de configuração (Tabela 7), seguindo o formato ARIMA(p, d, q).

Tabela 7: Combinações ARIMA(p, d, q) utilizadas

p	d	q
1	1	0
1	2	0
5	1	0
5	1	1
5	2	0
5	2	1
10	1	0

Devido as peculiaridades deste modelo, foi utilizado uma única informação dos dados pré-processados, valor de abertura (OPEN, Tabela 8), para executar o treino e posteriormente a previsão no mesmo período que a LSTM.

Tabela 8: Dados utilizados no modelo ARIMA

OPEN
313.823
314.234
312.998
305.005
296.396
⋮

Para gerar previsões dentro do período de validação dos dados, para cada configuração foi utilizado a metodologia de criar um modelo ARIMA após cada previsão, alimentando o modelo com o passo adquirido na iteração anterior. Ou seja, a cada iteração o conjunto de treinamento é utilizado para prever um valor e comparar com a validação desta iteração, posteriormente o treinamento é alimentado com o dado analisado na validação, repetindo este processo até o fim dos dados.

Ao final da etapa de previsões, é analisado os valores obtidos com os valores esperados através do ***Erro Quadrático Médio*** (**MSE**) obtendo valores para comparação dos diferentes estimadores. Todos os dados obtidos são apresentados no Capítulo 5.

4.4 Avaliação de Tendência

Apesar de fazer uso dos métodos supracitados na tentativa de predizer a tendência (*trending*), estes não diferem seu objetivo de tentar predizer o próximo valor de abertura (*OPEN*). Para prever a tendência em um determinado momento, utilizaremos o resultado das predições geradas para compor um sistema de avalição, onde é necessário comparar o dado predito (valor de abertura), com o dado anterior da série:

- $trending\_prediction(t) = OPEN_{t-1} - OPEN_t => x_t$
  - Retorna:
    - 1, se $x_t < 0$
    - 0, se $x_t = 0$
    - -1, se $x_t > 0$

Este sistema representa se uma determinada ação (*stock*) terá seu valor de abertura predito ($OPEN_t$), maior (1, se $x_t < 0$), igual (0, se $x_t = 0$) ou menor (-1, se $x_t > 0$) em relação ao seu valor anterior ($OPEN_{t-1}$). Os dados de previsão de tendência, independente do método utilizado, tem como valor inicial ($OPEN_{t-1}$) o valor real da última abertura do conjunto de treinamento, ou seja, todas as previsões de tendência partem do mesmo ponto inicial.

Tabela 9: Modelo de predição de tendência

t	$OPEN$	x
0	788.298	0
1	787.724	-1
2	799.825	1
3	799.825	0
4	811.881	1
⋮	⋮	⋮

Dessa forma, será utilizado o sistema descrito na avaliação do dado real e o dado estimado em uma matriz confusão, para que seja possível obter uma acurácia do método e seu erro quadrático médio.

5 RESULTADOS

Neste capítulo serão descritos os resultados obtidos das metodologias supracitadas no capítulo anterior (Capítulo 4). Todos os experimentos utilizaram os dados das séries temporais financeiras dos seguintes ativos, ***Amazon***, ***Ebay***, ***Facebook***, ***HP*** e ***Google***, e seu tratamento foi discutido anteriormente (Sessão 4.1).

Os resultados foram subdivididos pela metodologia e suas peculiariedades, inicialmente serão discutidas as **LSTMs** (Sessão 5.1) e sequentemente os resultados obtidos com o modelo **ARIMA** (Sessão 5.2). Posteriormente, os dados serão comparados e discutidos no Capítulo 5.

Para uma visualização dos dados mais assertiva, serão apresentadas os melhores valores obtidos com os métodos utilizados, segregando em duas tabelas distintas, contendo a configuração e *matriz confusão* e outra contendo a *acurácia* e o *MSE*. Dessa forma, todas os ativos (*stocks*) aparecem duas vezes, diferenciadas pelo asterisco em frente ao nome.

5.1 LSTM

Para este modelo foram criadas diferentes configurações na quantidade de unidades usadas na composição das redes, além de serem utilizadas diferentes tamanhos de *janelas*, demonstrados na Tabela 10.

Tabela 10: Configuração das LSTMs utilizadas

Unidades LSTM	Janelas de visualizações
[20, 30, 40, 50]	[1, 5, 10, 20, 24, 48, 72]

Como cada *janela* possui quantidades distintas de atributos, foram geradas 28 redes através das possíveis combinações, além de 28 redes semelhantes com diferença na presença de uma camada de ***dropout***, totalizando 56 configurações diferentes para cada ação. Dessa forma, temos 280 redes distintas avaliando 5 ações de diferentes empresas. Esta camada de dropout foi aplicada para comparar os efeitos de *overfitting* sobre os diferentes ativos.

Após a etapa de configuração das *LSTMs*, temos o tratamento dos dados, subdivididos em treinamento e validações. O conjunto de dados de treinamento foi consumido pelas diferentes configurações, demonstrados na sessão de apêndice. Posteriormente, com as *LSTMs* devidamente treinadas, os dados separados para as validações serão usados para metrificar o método.

Os resultados das matrizes confusões das diferentes configurações de **LSTMs**, são obtidos através da previsão da *tendência* do próximo valor de abertura (*OPEN*), 1 caso a previsão seja de aumento, 0 se for a mesma e -1 caso contrário.

Tabela 11: Melhores resultados das LSTMs

Stock	Unidades LSTM	Janela	Dropout	TN	FP	FN	TP
AMZN	30	1	True	69	25	39	118
EBAY	50	1	True	75	39	42	95
FB	30	1	True	73	36	38	104
HPQ	20	1	True	78	40	38	95
GOOGL	20	1	True	73	30	38	110
AMZN*	40	1	False	67	27	39	118
EBAY*	40	20	False	74	40	35	102
FB*	20	1	False	70	39	43	99
HPQ*	40	20	False	82	36	32	101
GOOGL*	50	1	False	71	32	38	110

Tabela 12: Acurácia e MSE das LSTMs

Stock	Acurracy	MSE
AMZN	0.745	1.019
EBAY	0.677	1.290
FB	0.705	1.179
HPQ	0.689	1.243
GOOGL	0.729	1.083
AMZN*	0.737	1.051
EBAY*	0.701	1.195
FB*	0.673	1.306
HPQ*	0.729	1.083
GOOGL*	0.721	1.115

Dentre as 280 configurações diferentes das *LSTMs*, podemos observar na Tabela 11 os melhores resultados da *matriz confusão*, entre configurações com a camada de *dropout* e sem essa camanda, posteriormente na Tabela 12 temos a *acurácia* e *MSE*.

A melhor predição da *tendência*, foi do ativo *AMZN-LSTM(30,1,True)* com 74.50% de *acurácia* e 1.0199 de *MSE*.

5.2 ARIMA

Para este modelo, foi utilizado o mesmo princípio do método anterior (*LSTM*), das configurações contidas na Tabela 7, combinadas a cada ativo (*stock*), foram obtidos os seguintes dados de previsão sobre os valores de abertura (*OPEN*) dentro do período de 2017, o mesmo período de previsões utilizado nas *LSTMs*.

Tabela 13: Melhores resultados utilizando ARIMA

Stock	p	d	q	TN	FP	FN	TP
AMZN	1	2	0	46	48	62	95
EBAY	1	2	0	44	70	76	61
FB	1	2	0	55	54	60	82
HPQ	1	2	0	48	70	58	75
GOOGL	5	2	1	49	54	56	92
AMZN*	10	1	0	39	55	58	99
EBAY*	5	1	1	39	75	73	64
FB*	5	2	0	49	60	63	79
HPQ*	10	1	0	53	65	64	69
GOOGL*	10	1	0	49	54	55	93

Tabela 14: Acurácia e MSE utilizando ARIMA

Stock	Acurracy	MSE
AMZN	0.561	1.752
EBAY	0.418	2.326
FB	0.545	1.816
HPQ	0.490	2.039
GOOGL	0.561	1.752
AMZN*	0.549	1.800
EBAY*	0.410	2.358
FB*	0.509	1.960
HPQ*	0.486	2.055
GOOGL*	0.565	1.737

Na Tabela 13, temos os melhores resultados aplicando as configurações descritas na mesma (*p*, *d* e *q*), dessa forma podemos observar os valores da *matriz confusão* (*True Negative*, *False Positive*, *False Negative* e *True Positive*).

Já na Tabela 14, temos o resultado da *acurácia* e *MSE* dos modelos descritos acima (Tabela 14), dentre eles, o melhor resultado é de 56.57% de *acurácia* e 1.7370 de *MSE* em *GOOGL-ARIMA(10,1,0)*.

5.3 LSTM x ARIMA

Nesta sessão iremos comparar a eficácia entre os dois médotos, apresentando os resultados por ativo na Tabela 15 e 16 sequentes.

Tabela 15: Acurácia: LSTM x ARIMA

Stock	LSTM	ARIMA
AMZN	0.745	0.561
EBAY	0.701	0.418
FB	0.705	0.545
HPQ	0.729	0.490
GOOGL	0.729	0.565

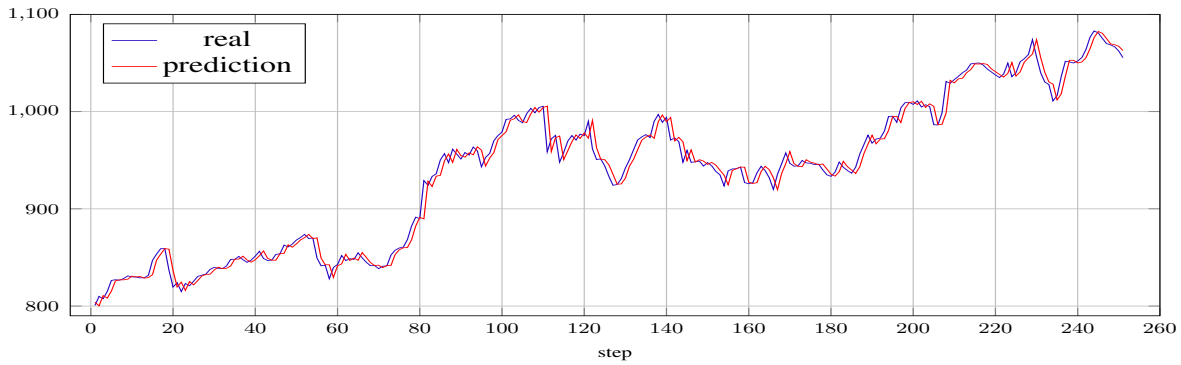
Tabela 16: MSE: LSTM x ARIMA

Stock	LSTM	ARIMA
AMZN	1.019	1.752
EBAY	1.195	2.326
FB	1.179	1.816
HPQ	1.083	2.039
GOOGL	1.083	1.737

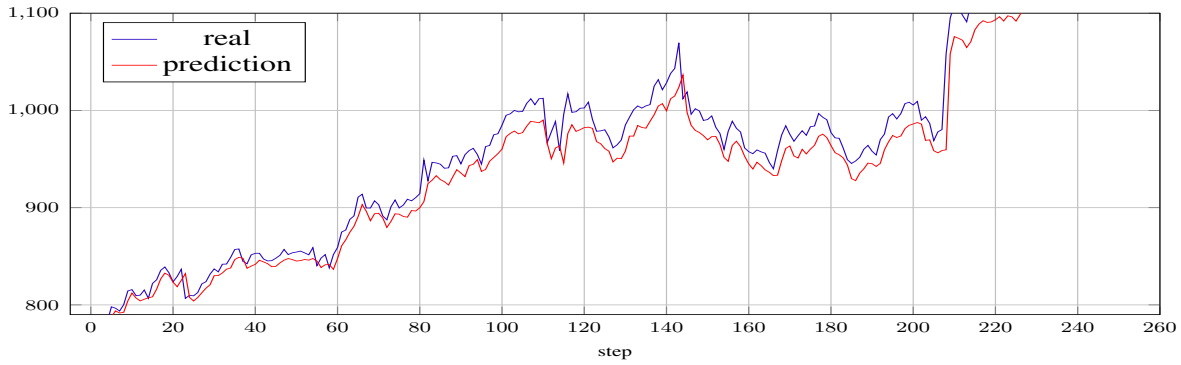
Podemos observar na Tabela 15 que em média, a avariação dos melhores resultados por ativo da *LSTM* ficam entre 74.5%-70.1% de *acurácia* e consequentemente, apresentando menor valor de *erro quadrático médio*. Já os melhores resultados por ativo do *ARIMA*, ficam na faixa de 56.5%-41.8% de *acurácia* e maior *erro quadrático médio*.

Evidentemente, o método *ARIMA* apresenta melhores resultados de previsão do real valor em comparação com as *LSTMs*, observado na Figura 9.

Apesar da melhor *acurácia* em determinar o próximo valor previsto utilizando *ARIMA*, podemos observar que o resultado de previsões da *LSTM* representa com maior grau de fidelidade a *tendência* do ativo.



(a) Previsão GOOGLE-ARIMA(10,1,0)



(b) Previsão AMAZON-LSTM-(1,30)

Figura 9: Comparação LSTM x ARIMA

6 CONCLUSÕES

Este trabalho teve o objetivo de propor um modelo de previsão de mercado financeiro através de uma *LSTM* alimentada unicamente com seus dados históricos e comparado com o modelo *ARIMA* para avaliar a *tendência de variação* do valor de abertura de um ativo.

A premissa é utilizar uma rede recorrente para verificar se haveria melhores previsões ao utilizar a memória de curto prazo para tentar identificar a variação futura (sinal) em relação ao valor corrente de abertura. Para esta finalidade foi utilizado diversas configurações de redes *LSTMs* sobre diferentes ativos e aplicando a mesma técnica utilizando o modelo *ARIMA* como base comparativa.

Com os resultados das previsões obtidas no capítulo anterior (Capítulo 5) sobre as *LSTMs*, verificamos que as diversas configurações tem diferenças significativas na capacidade de representar os dados previstos através de seus dados históricos, contudo, apesar de não se obter um resultado satisfatório em relação ao valor absoluto previsto e o real, o método se provou capaz de modelar com bom grau de *acurácia* a *tendência de variação* da série temporal.

Outro ponto importante a se notar é o tamanho da dimensionalidade de entrada que foi utilizado nos experimentos, que apesar de ser relativamente complexa dependendo da janela, não foi necessário utlizar nenhuma técnica de mudança de dimensão, como *feature selection*.

Podemos notar também, que a camada de dropout produz efeitos interessantes em determinados ativos, pela diferente variação temporal, peculiar de cada ativo, uma vez que seu comportamento é único.

Em relação ao modelo *ARIMA* utilizado, podemos verificar que ele é bem estável e apresenta resultados satisfatórios como *baseline* comparativa para valores previstos. Mas ao aplicar o modelo para avaliar a *tendência de variação*, o método se torna menos eficaz, com uma taxa de *acurácia* máxima de 0.5657370517928287 para o ativo **GOOGL** (*Google*) e menores valores para os demais.

6.1 Contribuições

Neste trabalho foi idealizado uma nova proposta de avaliação do sinal futuro do valor de abertura de um determinado ativo (*tendência de variação*), utilizando somente seus dados históricos como parâmetros de entrada através de uma rede recorrente.

Consequentemente, para validar o modelo proposto foram realizado experimentos de desempenho e comparados com outras técnicas existentes (*ARIMA*).

Dessa forma podemos notar que através do uso da memória de curto prazo, a *LSTM* foi capaz de modelar com certo grau de acurácia a *tendência de variação* do valor de abertura futura, gerando indícios que deve haver correlação em seus dados históricos e futuros, contrariando a

hipótese de *Mercado Eficiente*.

6.2 Trabalhos Futuros

Como a área de estudos financeiros e previsões da bolsa de valores é muito visada, este campo provê uma infinidade de caminhos a serem tomados para extender este trabalho.

Para melhorar a qualidade das previsões, temos alguns caminhos não excludentes a serem considerados, como relacionar os dados históricos com variáveis exógenas para tentar aumentar a *acurácia* do método, uma vez que através das diferentes configurações das *LSTMs* para os diferentes ativos, não foi obtido ganhos significativos. Dessa forma existe a hipótese de que para melhorar a previsão existem dados exógenos não levados em consideração. Outro caminho seria trabalhar na construção da *LSTM*, utilizando redes profundas e diferentes funções de ativação ou sobreposição de redes (**redes convolucionais**).

Devemos levar em consideração a janela de previsões gerada através das *LSTMs*, uma vez que através dos dados históricos para treinamento, período de 2010 até 2016, foram geradas janelas de 1 ano de previsão. Ao contrário do método *ARIMA* utilizado para conparações, que somente faz previsões de um dia após o treinamento. Neste caso, há a hipótese de que um período de previsão menor poderia gerar resultados mais acurados.

Outro ponto a ser observado e trabalhado no futuro é de quantificar a taxa de variação da tendência para uma melhor assertividade, obtendo informações sobre a real *tendência* de mercado.

REFERÊNCIAS

CORTEZ, B. et al. An architecture for emergency event prediction using LSTM recurrent neural networks. *Expert Systems with Applications*, Elsevier Ltd, v. 97, p. 315–324, 2018. ISSN 09574174. Disponível em: <https://doi.org/10.1016/j.eswa.2017.12.037>. Citado na página 36.

FAMA, E. F.; MALKIEL, B. G. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, v. 25, p. 383–417, 1970. ISSN 1540-6261. Citado na página 35.

FARIA, E. L. D. et al. Expert Systems with Applications Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods. *Expert Systems With Applications*, Elsevier Ltd, v. 36, n. 10, p. 12506–12509, 2009. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2009.04.032>. Citado 2 vezes nas páginas 36 e 37.

GURESEN, E.; KAYAKUTLU, G.; DAIM, T. U. Expert Systems with Applications Using artificial neural network models in stock market index prediction. *Expert Systems With Applications*, Elsevier Ltd, v. 38, n. 8, p. 10389–10397, 2011. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2011.02.068>. Citado na página 36.

HASSAN, J. ARIMA and regression models for prediction of daily and monthly clearness index. *Renewable Energy*, Elsevier Ltd, v. 68, p. 421–427, 2014. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2014.02.016>. Citado na página 36.

HEDAYATI, A.; HEDAYATI, M.; ESFANDYARI, M. Stock market index prediction using artificial neural network. v. 21, p. 89–93, 2016. Citado na página 36.

KIM, K.-j. Financial time series forecasting using support vector machines. v. 55, p. 307–319, 2003. Citado na página 35.

MALKIEL, B. G. A Random Walk Down Wall Street. 1973. Citado na página 35.

MCNALLY, S.; ROCHE, J.; CATON, S. Predicting the Price of Bitcoin Using Machine Learning. *Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018*, p. 339–343, 2018. ISSN 2015/2016. Citado na página 36.

QIU, M.; SONG, Y. Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. p. 1–11, 2016. Citado na página 35.

QIU, M.; SONG, Y.; AKAGI, F. Application of artificial neural network for the prediction of stock market returns : The case of the Japanese stock market. Elsevier Ltd, v. 85, p. 1–7, 2016. Citado na página 37.

SIAMI-NAMINI, S.; NAMIN, A. S. Forecasting Economics and Financial Time Series: ARIMA vs. LSTM. p. 1–19, 2018. Disponível em: <http://arxiv.org/abs/1803.06386>. Citado na página 36.

TICKNOR, J. L. Expert Systems with Applications A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems With Applications*, Elsevier Ltd, v. 40, n. 14, p. 5501–5506, 2013. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2013.04.013>. Citado na página 37.

TIME SERIES. In: WIKIPÉDIA: a enciclopédia livre. Wikimedia, 2018. Disponível em: <https://en.wikipedia.org/wiki/Time_series>. Acesso em: 01 Jun. 2018. Citado na página 26.

XIAOGUO, W.; YUEJING, L. Arima time series application to employment forecasting. *Proceedings of 2009 4th International Conference on Computer Science and Education, ICCSE 2009*, p. 1124–1127, 2009. Citado na página 36.

YANG, R. et al. Simulation of rain attenuation time series by ARIMA model. *2013 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, CSQRWC 2013*, n. 4, p. 304–307, 2013. Citado na página 36.

ZHANG, J.; NAWATA, K. A comparative study on predicting influenza outbreaks. *BioScience Trends*, v. 11, n. 5, p. 533–541, 2017. ISSN 18817823. Citado na página 37.

ZHE, S.; YU, J.; ZIJUN, Z. Short-term wind speed forecasting with Markov-switching model. *Applied Energy*, v. 130, p. 103–112, 2014. ISSN 03062619. Disponível em: <http://search.ebscohost.com.proxy-ub.rug.nl/login.aspx?direct=true&db=aph&AN=97191733&site=ehost-live&scope=site>. Citado na página 36.