UNIVERSIDADE JÚLIO MESQUITA FILHO
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE ALGORÍTMOS DE REGRESSÃO APLICADOS AO MERCADO FINANCEIRO

DALTON AKIO OKAMURA

Bauru 2019

DALTON AKIO OKAMURA

ANÁLISE DE ALGORÍTIMOS DE REGRESSÃO APLICADOS A MERCADO FINANCEIRO

Trabalho de Conclusão de Curso apresentado junto ao Curso de Bacharelado de Ciência da Computação da Universidade Júlio Mesquita Filho, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador:

Professora Dra. Simone Domingues Prado

Bauru, 2019

DALTON AKIO OKAMURA

ANÁLISE DE ALGORÍTMOS DE REGRESSÃO APLICADOS AO MERCADO FINANCEIRO

Trabalho de Conclusão de Curso apresentado junto ao Curso de Bacharelado de Ciência da Computação da Universidade Júlio Mesquita Filho, como requisito parcial à obtenção do título de Cientista da Computação.

Aprovada em 12/06/2019.

Doutora Simone das Graças Domingues Prado - Orientador
Departamento de Computação - Faculdade de Ciências - Unesp/Bauru

Doutora Kleber Rocha de Oliveira - Filiação
Campus Experimental de Rosana - Engenharia de Energia - Unesp/Rosana

Kelton Augusto Pontara da Costa
Departamento de Computação - Faculdade de Ciências - Unesp/Bauru

Este trabalho é dedicado aos meus pais, que batalharam a vida por mim e sem eles não estaria onde estou.

AGRADECIMENTOS

Agradecimento aos meus pais, por todos os seus sacrifícios e esforços, sem eles eu não seria ninguém. Gostaria também de agradecer a minha orientadora, professora Simone Domingues Prado, que me ajudou durante toda minha trajetória da graduação.

“O fator decisivo para vencer o maior obstáculo é, invariavelmente, ultrapassar o obstáculo anterior.”

Henry Ford

RESUMO

Em um momento onde a visibilidade e popularidade do mercado financeiro esta cada vez maior, assim como o seu número de investidores e existem grandes riscos ligados a este mundo, que podem ocasionar até a perda total do patrimônio de seus especuladores. O trabalho aqui desenvolvido visa promover a interdisciplinaridade entre a computação e a economia, com o intuito de fornecer conhecimento aos interessados dos dois mundos. Serão apresentados conceitos da psicologia dos investidores, análise técnica, que é a ciência estatística que objetiva prever o valor dos ativos, e conceitos de *machine learning*. Baseando-se nesses conceitos buscou-se compreender quais parâmetros possuem maior peso dentro dos modelos de regressão. Testes foram realizados em algorítmos, RLM, SVR e RFR.

Palavras chaves: **mercado financeiro, economia, análise técnica, regressão.**

ABSTRACT

We live in a moment where the visibility and popularity of the financial market is increasing, as is its number of investors. Despite their potential for monetary return, there are large risks linked to this world, which can even lead to the total loss of the equity of their speculators. With this in mind, the work proposed here aims to promote the interdisciplinarity between computing and economics, in order to provide knowledge to the stakeholders of both worlds. It will be presented concepts of investor psychology, technical analysis, which is the statistical science that aims to predict the value of the assets, and machine learning concepts to understand which parameters have the greatest weight within the regression models.

LISTA DE FIGURAS

Figura 1 – Gráfico Ibovespa 21
Figura 2 – Fórmula *Stochastic Oscilator* 22
Figura 3 – Fórmula *Money Flow Index* 23
Figura 4 – *Relative Strength Index* 24
Figura 5 – Fórmula Relative Strength Index 24
Figura 6 – Fórmula Média Móvel Exponencial 25
Figura 7 – Gráfico MACD 26
Figura 8 – Fórmula regressão linear 28
Figura 9 – Algorítimo de *Backwards Elimination* 30
Figura 10 – Representação do SVM no hiperplano 31
Figura 11 – Problema primal 32
Figura 12 – Problema dual 32
Figura 13 – Função de decisão 32
Figura 14 – Representação do processo de *machine learning* 37
Figura 15 – Resultado da regressão vs Valor real 39
Figura 16 – Resultado após aplicação do *Backward Elimination* 40
Figura 17 – Resultado da regressão vs Valor real SVM 41
Figura 18 – Resultado da regressão vs Valor real RFR 42
Figura 19 – ML vs SVR vs RFR 45

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
TCC	Trabalho de Conclusão do Curso
NBR	Norma Brasileira
VOC	Vereennigde NederlandscheOostindische Compagnie
IPO	Inicial Public Offering
ML	Machine Learning
LEI	Leading Indicators
LAI	Lagging Indicators
NB	Non-bounded
RSI	Relative Strength Index
SO	Stochastic Oscilator
MFI	Money Flow Index
MACD	Moving Average Convergence/Divergence
LS-SVM	Least Square Support Vector Machine
PSO	Particle Swarm Optimization
SVR	Support Vector Regression
MME	Média Móvel Expónencial
IDE	Integral Development Enviroment
RLS	Regressão Linear Simples
SVM	Support Vector Machine
OOB	Out of Bag
RFR	Random Forest Regression

1 INTRODUÇÃO 14

1.0.1 Objetivos Gerais 15

1.0.1.1 Objetivos Específicos 15

1.0.2 Problema 15

1.0.3 Justificativa 16

2 MERCADO FINANCEIRO 17

3 ECONOMIA COMPORTAMENTAL 19

4 ANÁLISE TÉCNICA 20

4.1 Indicadores Técnicos 22

4.1.1 *Stochastic Oscillator* (SO) 22

4.1.2 *Money Flow Index (MFI)* 22

4.1.3 Relative Strength Index (RSI) 23

4.1.4 Moving Average Convergence/Divergence (MACD) 25

5 *MACHINE LEARNING* (ML) 27

5.1 Regressão Linear Simples (RLS) 28

5.2 Algorítimo de Regressão Multivariável 29

5.3 Support Vector Regression (SVR) 30

5.3.1 Funcionamento Básico 31

5.3.2 Formulação Matemática 31

5.4 *Random Forest Regression* (RFR) 32

5.4.1 Algorítimo 33

6 METODOLOGIA DE PESQUISA 34

6.1 Ferramentas Utilizadas 34

6.1.1 *Python* 34

6.1.2 *Scikit-learn* 34

6.1.3 Pandas 35

6.1.4 Anaconda 35
6.1.5 *Spyder* 35
6.1.6 *Overleaf* 35
7 DESENVOLVIMENTO 36
7.1 *Dataset* 36
7.2 *Data PreProcessing* 36
7.3 Algorítimos 38
7.3.1 Aplicar o algorítimo ao modelo candidato 38
8 RESULTADOS 39
8.1 Regressão Linear Múltipla aplicada à Ambev 39
8.2 SVR aplicado à Ambev 40
8.3 RFR aplicado à Ambev 42
9 CONCLUSÃO 44
REFERÊNCIAS 46

1 INTRODUÇÃO

A cada dia que passa o protagonismo do mercado financeiro aumenta no cenário econômico brasileiro. A era da informação trouxe diversos benefícios para nossa geração, porém em um contexto de investimento é necessário entender quais delas são relevantes e quais devem ser desconsideradas. Tendo em vista os riscos e benefícios de se investir no mercado de ativos, o trabalho aqui proposto visa fornecer conhecimento científico aos interessados do mundo financeiro e da computação, para que estes estejam munidos do maior número de ferramentas possíveis, para criarem suas próprias estratégias de investimento.

Para possuir o pleno entendimento do mercado financeiro, é necessário conhecer o contexto histórico que condicionou seu surgimento, pois assim o investidor leigo é capaz de compreender quais as necessidades levaram a sua criação e como ele evoluiu até se tornar o que é hoje. Este tópico é abordado no Capítulo 2.

Após entender um pouco das condições que ocasionaram o surgimento do mercado, é necessário compreender a mentalidade daqueles que o constituem nos dias de hoje, pois este conhecimento fornece noções ao investidor de como montar estratégias efetivas que maximize o lucro e minimize as perdas. Tal assunto é abordado Capítulo 3.

Após os conceitos apresentados na capítulos 2 e 3, o leitor estará munido de conhecimento suficiente para entender a análise técnica, que é a ciência estatística que busca prever o preço das ações tendo em vista a psicologia dos investidores. O capítulo 4 cobre este assuntos, e apresenta os 5 parâmetros técnicos que serão utilizados nos modelos de *machine learning* implementados neste trabalho.

Após a apresentação de todo o conteúdo de economia dos capítulos 2 ao 4, no capítulo 5 discute conceitos de *machine learning* e os modelos que serão implementados no trabalho aqui proposto.

O capítulo 6 descreve a metodologia utilizada na implementação dos modelos de *machine learning*.

O capítulo 7 descreve o processo de desenvolvimento realizado que gerou os resultados apresentados no capítulo 8.

1.0.1 Objetivos Gerais

Este trabalho propõe o fornecimento de uma base empírica aos investidores do mercado financeiro que buscam maximizar seus lucros e minimizar suas perdas. Para tal, serão apresentados conceitos sobre economia, computação e matemática estatística que são bases fundamentais para a criação de estratégias efetivas de mercado.

1.0.1.1 Objetivos Específicos

- Modelar e aplicar conceitos de *machine learning* em mercado financeiro.

- Prever através de modelos de regressão o preço de ativos brasileiros.

- Promover interdisciplinaridade.

1.0.2 Problema

Até o século XX as bases empíricas seguidas pelos grandes estudiosos partiam do pressuposto de que o humano é um ser racional, e assim concluíram que o mercado seguia um padrão lógico. Recentemente surgiram novos estudos como a economia comportamental (THALER, 2011), que teve como um dos grandes pilares o psicólogo Richard H. Thaler, de 72 anos. Thaler recebeu em 2017 o Prêmio Nobel de Economia, devido as suas pesquisas que aplicam conceitos de psicologia ao paradigma econômico.

(THALER, 2011) comprovou em seus estudos o teorema de que os indivíduos têm maior dificuldade em abrir mão de um benefício no presente do que de um benefício no futuro. Essa teoria é comprovada ao observar o comportamento humano na compra de ativos. As pessoas veem o preço de uma ação subir, e compram quando seu preço está em alta, pois existe a expectativa de que o valor do ativo continuará a subir. O mercado de ações é especulativo, fator que causa a irracionalidade humana, pois em qualquer outro mercado a estratégia racional seria comprar o produto a um preço baixo e revendê-lo por um superior e não comprá-lo por um valor elevado com a esperança de que seu valor continue a aumentar.

Com a teoria de *Thaler* pode-se observar que o instinto humano se baseia na busca de conquistas em prazos curtos. *Warren Buffett* um dos maiores investidores da atualidade e grande ícone da economia mundial possui diversas frases como ”Não importa quão grande o talento ou esforço, algumas coisas levam tempo”, “Nosso período

favorito de manutenção de uma ação é para sempre”, (CASTRO, 2019). A teoria de (THALER, 2011) nas palavras de *Buffett*, no que diz respeito a falta de racionalidade de parte dos investidores. Tendo o fator humano como a problemática do mercado financeiro, o trabalho desenvolvido tem por finalidade fornecer conhecimento para que as pessoas possam tomar decisões racionais dentro de operações mercadológicas e não sejam ofuscadas por seus instintos.

1.0.3 Justificativa

Através da economia comportamental, foi verificada a existência de um grande número de investidores que realizam seus investimentos baseados em emoções. O trabalho aqui desenvolvido objetiva aumentar e promover a racionalidade do mercado, para maximizar lucros e minimizar perdas, fato que possibilita à população possuir um renda extra ao investir no mercado financeiro (THALER, 2011).

2 MERCADO FINANCEIRO

A história da bolsa de valores se inicia quando em 1596 o navegador holandês *Jan Huygen van Linschote* que participou de diversas expedições em navios portugueses, lançou o livro "Relato de uma viagem pelas navegações dos portugueses no Oriente". Nele, constavam um grande número de informações e técnicas marítimas acumuladas pelos portugueses (VERSIGNASSI, 2009). Em 1596 a Holanda era uma força emergente, porém estava atrás das duas grandes potências da época: Portugal e Espanha. Com o desejo de acompanhar a expansão marítima que dominou o período dos séculos XV ao século XVIII a Holanda necessitava de dinheiro para financiar suas navegações rumo ao Oriente. Para pagar suas expedições, as seis Companhias das Índias da época se juntaram em uma megacorporação: *Vereennigde Nederlandsche Oostindische Compagnie* (VOC).

A grande ideia da VOC para arrecadação de investimentos foi "Convidar a população para virar sócia" (VERSIGNASSI, 2009), ideologia que culminou no surgimento da primeira bolsa de valores da História: *Amsterdam Stock Exchange*.

A estratégia inventada pela VOC no final do século XVI representa o conceito de *Inicial Public Offering* (IPO), em português Oferta Pública Inicial, termo extremamente conhecido na atualidade. Diversas empresas como *Microsoft, Google, Ambev, Apple* entre tantas outras gigantes tornaram o seu capital público para levantar dinheiro dos investidores para concretizar seus planos de expansão. No processo de uma IPO a empresa se divide em milhares de partes, onde cada parte é a representação de uma ação. Depois é feito uma avaliação de quanto a companhia vale e este valor é dividido pelo número total das ações. Esse mecanismo serve para que grandes empresas levantem capital para expansão de suas operações. (JAMES, 2019b)

O mercado financeiro é um termo que se aplica à qualquer mercado onde há a negociação de títulos, ações, moedas, derivativos (contratos que possuem seu valor derivado de um ativo subjacente, taxa de referência ou índice) e *commodities*. (KENTON, 2019). O preço de uma ação é o simples reflexo da quantidade de oferta e demanda que existe por ela, se existem muitos compradores seu valor sobe, porém caso a quantidade de vendedores prevaleça o seu valor decresce.

Apesar de parecer simples este mecanismo baseado em oferta e demanda é o coração da definição do mercado livre. O valor de um ativo é o reflexo de quanto a empresa vale aos olhos dos investidores, portanto as companhias com maior lucro,

boas taxas de dividendos e com diversas outras características consideradas atrativas, fazem com que o valor de um ativo decole. Porém fechamentos negativos, notícias que sugerem uma instabilidade financeira, administrativa ou corrupção podem levar a queda dos preços. Através destes fatos, pode-se averiguar que o valor de uma ação nem sempre é o reflexo de um fator macroeconômico e sim de especulações da saúde financeira de uma empresa.

3 ECONOMIA COMPORTAMENTAL

A economia comportamental busca entender o processo de decisões econômicas, tanto de indivíduos quanto de instituições (KENTON, 2017).

No mundo tradicional da economia existe a hipótese de que o humano é um ser racional, que não se deixa levar pelas emoções e possui capacidade de avaliar os diversos contextos econômicos para ponderar qual a melhor decisão a ser tomada (THALER, 2011).

Em contra partida a economia comportamental afirma que os humanos são seres irracionais e que grande parte de suas escolhas são reflexos de suas emoções (KENTON, 2017). Como são pessoas que ditam o comportamento do mercado, é necessário entender o mecanismo de tomada de decisão das massas e como esse conhecimento é explorado pelas grandes corporações.

O ser humano possui diversos comportamentos automáticos que são desenvolvidos ao longo do tempo através das influências sociais que são submetidos no decorrer da vida (CIALDINI, 2009). No livro de Cialdini, são apresentados inúmeros conceitos sobre o comportamento dos indivíduos e como eles são exploradas pelas grandes corporações, para aumentar sua efetividade no mercado.

O princípio do contraste é amplamente abordado no primeiro capítulo de sua obra. O cérebro humano funciona através de comparações e quando são apresentados a dois conceitos que são opostos, perde-se a real noção do que cada um representa. Uma estratégia utilizada por corretores de imóveis é apresentar inicialmente ao cliente um imóvel que esteja em péssimas condições, para que o segundo imóvel (em boas condições) possua maior atratividade. Mesmo que o segundo local apresentado estivesse longe do ideal, o princípio do contraste acaba tornando-o mais atrativo do que ele realmente é. (CIALDINI, 2009)

Entender os mecanismos psicológicos que estão por trás do comportamento em massa de investidores é fundamental, pois tal entendimento auxilia no refinamento das estratégias para torna-las mais efetivas.

4 ANÁLISE TÉCNICA

A análise técnica é o estudo do histórico dos dados das ações e inclui variáveis como o preço de abertura, fechamento, máximos, mínimos e o volume do ativo. Seu grande objetivo é determinar se a tendência do mercado irá se manter ou reverter, baseando-se no seu comportamento prévio, buscando uma oportunidade de compra em tendências de alta e venda na de baixa (JAMES, 2019b). São utilizados conceitos de economia comportamental (Capítulo 3) e análise quantitativa, que é o estudo do comportamento através de modelos matemáticos estatísticos. A aplicação da análise técnica, é feita através do uso de padrões gráficos e indicadores técnicos. Através da análise técnica é possível avaliar se um ativo esta sobre-comprado (super-valorizado) ou sobre-vendido (desvalorizado).

Apesar de existirem diversos algorítimos de *machine learning* (ML) focados na análise gráfica, o trabalho aqui proposto se utiliza da análise quantitativa, através do uso de indicadores técnicos. Enquanto a análise gráfica gera informações que são subjetivas para a tomada de decisão, os indicadores técnicos são representações estatísticas que servem como dados para confirmar as conclusões tiradas dos padrões gráficos. (KUEPPER, 2017)

Figura 1 – Gráfico Ibovespa

Fonte: Disponível em: <https://br.investing.com/indices/bovespa>. Acesso em: 20 de abril de 2019.

Os indicadores técnicos podem ser divididos principalmente em duas categorias: Os *Leading Indicators* (LEI) e *Lagging indicators* (LAI). Os LEI precedem os movimentos dos preços, buscam prever o futuro e possuem maior efetividade em contextos de mercado onde não há uma tendência clara, ou em períodos de consolidação (termo utilizado para indicar que o preço do ativo está estável) da ação. Já os LAI seguem o movimento dos preços e são utilizados como ferramentas de confirmação. Em circunstâncias em que o mercado esta claramente em alta ou em baixa, os LAI indicam a força de continuação ou reversão da tendência em questão. (KUEPPER, 2017)

Os indicadores também poder ser classificados em *Oscillator* e os *Non-bounded* (NB). O *Oscilator* é o tipo mais comum de indicador e trabalha dentro de

um limite entre 0 e 100, onde valores próximos de 0 indicam uma condição de sobre venda e 100 de sobre compra. Já os NB são minoria dentro da análise técnica e tentam indicar a força da tendência do mercado sem a utilização dos limites.

São utilizados quatro indicadores da análise técnica como variáveis independentes dos algorítimos de *machine learning*, sendo eles: *Relative Strength Index* (RSI), *Money Flow Index* (MFI), *Stochastic Oscillator* (SO) e *Moving Average Convergence/Divergence* (MACD), que foram utilizados por (HEGAZY, 2013) em seus experimentos. Neste artigo os pesquisadores implementaram o *Least Square Support Vector Machine* (LS-SVM) para predição de preços dos ativos, e os parâmetros de entrada do LS-SVM foram otimizados através do uso do algorítimo meta-heurístico *Particle Swarm Optimization* (PSO). Todos estes conceitos são discutidos nos capítulos posteriores.

4.1 INDICADORES TÉCNICOS

4.1.1 *Stochastic Oscillator* (SO)

O SO é um indicador técnico da categoria dos osciladores que compara o fechamento de um ativo específico com seus valores anteriores em um determinado período de tempo. O período mais utilizado é o de 14 dias. Ele serve para identificar condições de sobre compra e sobre venda, e oscila entre os valores de 0 a 100. Um SO de 80 para cima sugere um mercado de sobre compra, já abaixo de 20 sobre venda. A Figura 2 representa a fórmula do SO.

Figura 2 – Fórmula *Stochastic Oscilator*

```
%K = (Current Close - Lowest Low)/(Highest High - Lowest Low) * 100
%D = 3-day SMA of %K

Lowest Low = lowest low for the look-back period
Highest High = highest high for the look-back period
%K is multiplied by 100 to move the decimal point two places
```

Fonte: Disponível em: <https://br.investing.com/indices/bovespa>. Acesso em: 20 de abril de 2019.

4.1.2 Money Flow Index (MFI)

O MFI é calculado através da média do preço de fechamento do ativo, seus valores mais altos e baixos multiplicando o resultado pelo volume diário das ações negociadas naquele dia. (MITCHELL, 2019)

O MFI é utilizado para confirmar tendências das forças compradoras evendedoras. Ao comparar o valor do indicador no dia anterior com o atual, o analista sabe se o fluxo de dinheiro foi positivo ou negativo. Um MFI atual maior que o dia anterior sugere que pressão compradora esta aumentando, enquanto um menor indica que a força vendedora possui vantagem. (MITCHELL, 2019)

O *Money Flow Index* (MFI) é um indicador da categoria dos osciladores e se utiliza do MFI para realização dos seus cálculos. MFI acima de 80 sugere um mercado com a força compradora predominante e abaixo de 20 indica vantagem para a força vendedora. A figura 3 demonstra a fórmula do MFI. (MITCHELL, 2019)

Figura 3 – Fórmula *Money Flow Index*

The Formulas for the Money Flow Index (MFI) are

$$\text{Money Flow Index} = 100 - \frac{100}{1 + \text{Money Flow Ratio}}$$

Where:

$$\text{Money Flow Ratio} = \frac{\text{14 Period Positive Money Flow}}{\text{14 Period Negative Money Flow}}$$

$$\text{Raw Money Flow} = \text{Typical Price} * \text{Volume}$$

$$\text{Typical Price} = \frac{(\text{High} + \text{Low} + \text{Close})}{3}$$

Fonte: Disponível em: <https://br.investing.com/indices/bovespa>. Acesso em: 20 de abril de 2019.

4.1.3 Relative Strength Index (RSI)

O indicador técnico RSI foi criado por *Welles Wilder Junior* no livro *New Concepts in Technical Trading Systems* em 1978. Ele é um indicador da classe dos osciladores e mede a mudança recente do valor dos ativos, para avaliar se situações de sobre compra (super valorizado) ou sobre venda (sub valorizado). O RSI é demonstrado em um *range* entre 0 e 100, onde valores acima de 70 indicam que o preço do ativo pode estar super valorizado, enquanto valores abaixo de 30 indicam uma desvalorização. A Figura 4 representa a fórmula para realização do cálculo do RSI.(JAMES, 2019a)

Figura 4 – *Relative Strength Index*

$$RSI = 100 - \frac{100}{1 + RS}$$

$$RS = \text{Average Gain} / \text{Average Loss}$$

Fonte: Disponível em: <https://www.investopedia.com/terms/r/rsi.asp>. Acesso em: 05 de abril de 2019.

O RSI sobe conforme ocorrem fechamentos positivos do mercado e desce conforme os negativos. A Figura 5 representa os dados do oscilador juntamente a representação gráfica em forma de *candles* do ativo.

Figura 5 – Fórmula Relative Strength Index

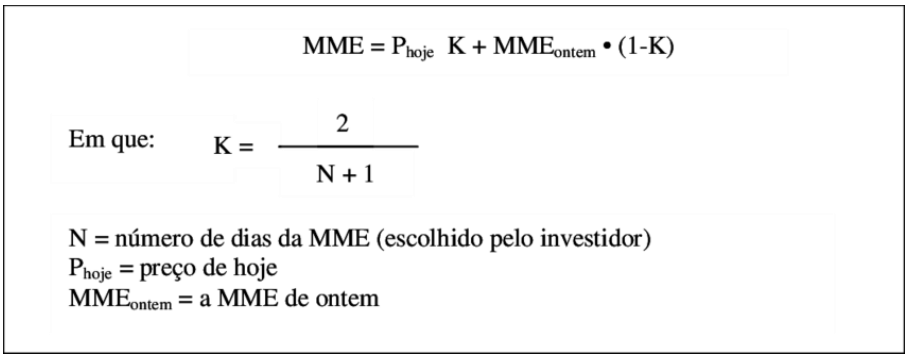
Fonte: Disponível em: <https://www.investopedia.com/terms/r/rsi.asp>. Acesso em: 05 de abril de 2019.

4.1.4 Moving Average Convergence/Divergence (MACD)

O MACD é o indicador técnico mais utilizado da análise técnica (RADAR, 2018), por sua característica de rápida detecção de fortes tendências de curto prazo e se baseia no conceito de Média Móvel. A Média Móvel tem por objetivo prever a direção dos preços do ativo com um certo nível de atraso, pois seu cálculo se baseia em dados passados. A Média Móvel Exponencial (MME) é calculada através da fórmula representada pela Figura 6. A Figura 7 mostra um gráfico onde a linha azul representa um MME de 12 dias e a vermelha de 26. Ao analisar o gráfico, pode-se interpretar que no momento em que a linha azul ultrapassa a vermelha existe uma alta recente dos preços que supera o crescimento anterior. O MACD se baseia na diferença entre os MMEs, mais especificamente, subtrai o MME de período mais longo pelo curto. Esses *timeframes* podem variar de acordo com a estratégia do investidor, mas é comum o uso de períodos de 12 a 26 dias. (RADAR, 2018)

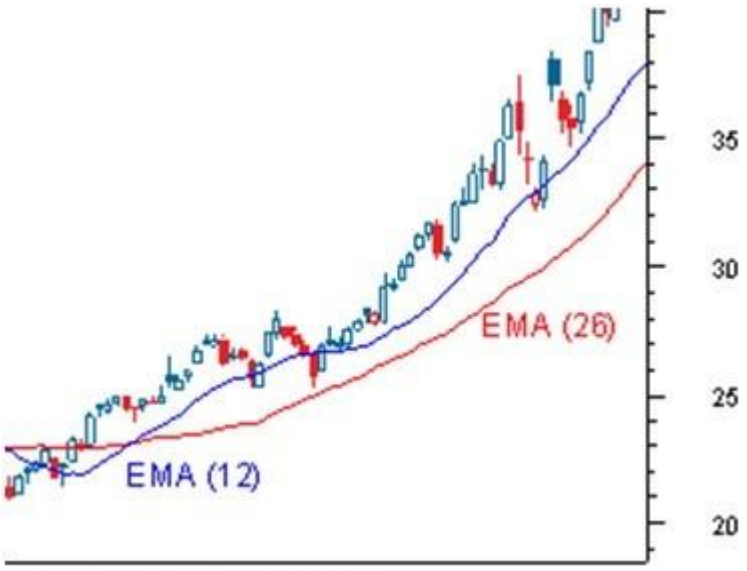
Quando o MME de curto prazo esta acima do longo, pode-se identificar uma possível tendência altista. Se houver confirmação dessa análise através da observação de outros indicadores, existe a possibilidade de compra para um mercado de alta. Já quando o valor do MME de períodos maiores é superior, há indicações de um mercado de baixa, podendo sinalizar uma opção de venda (RADAR, 2018).

Figura 6 – Fórmula Média Móvel Exponencial



Fonte: Disponível em: <https://www.tororadar.com.br/investimento/analise-tecnica/macd>. Acesso em: 05 de abril de 2019.

Figura 7 – Gráfico MACD



Fonte: Disponível em: <https://www.tororadar.com.br/investimento/analise-tecnica/macd>. Acesso em: 05 de abril de 2019.

5 *MACHINE LEARNING* (ML)

O ML utiliza o princípio de inferência chamado indução, que objetiva analisar um domínio específico para obtenção de conclusões genéricas respectivas ao conjunto analisado. O aprendizado indutivo pode ser dividido em supervisionado e não supervisionado. (LORENA, 2007)

O aprendizado supervisionado pode ser entendido através da figura de um professor externo, que apresenta um domínio de entradas com suas respectivas saídas. Assim o algorítimo objetiva aprender a relação entre o conjunto de entrada e saída, para reproduzir este conhecimento a dados desconhecidos. (LORENA, 2007)

Já no aprendizado não supervisionado, não existe a figura do professor que apresenta um conjunto de exemplos do domínio. Em vez disso, o algorítimo busca agrupar as entradas de dados submetidas de acordo com uma medida de qualidade. Essa metodologia é utilizada para identificar padrões dos dados ou tendências que ajudem na interpretação dos dados. (LORENA, 2007)

Quando os rótulos ou classes de representação do fenômeno assumem valores discretos de 1 a k, caracteriza-se um problema de classificação. Caso os rótulos estejam na ordem de valores contínuos, o problema passa a ser de regressão. A diferença primordial entre classificação e regressão são as suas finalidades, onde o classificador analisa as características e atribui os dados a um domínio conhecido, enquanto as regressões buscam realizar previsões através dos dados obtidos previamente. (LORENA, 2007)

Os problemas de ML são modelados através do uso de vetores, que agrupam variáveis independentes ou características x do problema, que em conjunto podem predizer ou classificar o valor da variável dependente y. Existem duas categorias para estes atributos que compõe o problema, sendo eles os atributos nominais e os contínuos. (LORENA, 2007)

Atributos nominais são aqueles dos quais não possuem valor quantitativo, e possuem apenas características categóricas, tais como cor, raça ou região. Já nos atributos contínuos é possível encontrar uma ordem quantitativa dos valores assumidos,tais como no peso, altura, lucro, investimento. (LORENA, 2007).

Para o trabalho aqui proposto, são utilizados algorítimos de ML focados para problemas de regressão, cujas características ou atributos independentes são os

indicadores técnicos apresentados no Capítulo 4. Neste capítulo são discutidos os algorítimos de RLS e o *suport vector regression*, que são utilizados na aplicação aqui proposta. Ambos possuem modelagens para resolver problemas de classificação e regressão, e o foco deste trabalho é baseado na regressão.

Um aspecto importante dos algorítimos de ML são os ruídos. Esses ruídos podem aparecer em formas de dados, atributos ou rótulos que possuem interpretações fora do domínio comum dos outros dados. Além dos ruídos, existem também os *outliers*, que são exemplos extremamente diferentes dos ordinários presentes dentro do dataset, mas que também fazem parte do domínio da solução. Esses casos devem ser tratados e minimizados para não moldar e diminuir a assertividade do algorítimo de aprendizado. É importante possuir um conjunto de dados coerente com o domínio desejado, para que diminuir a possibilidade de predições ou classificações incoerentes. (LORENA, 2007)

5.1 REGRESSÃO LINEAR SIMPLES (RLS)

Algorítimos de regressão linear simples são uma das abordagens mais básicas de aprendizado supervisionado. Apesar de sua simplicidade, são extremamente efetivos e resolvem de forma satisfatória problemas de baixo grau de dificuldade. A palavra linear sugere que a relação entre o parâmetro dependente e o independente pode ser descrito através de uma reta, ou algo bem próxima a isso, seguindo a fórmula representada na figura 8. (MENON, 2017a)

Figura 8 – Fórmula regressão linear

$$y = ax + b \qquad a = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Fonte: Feita pelo autor

A figura 8 é a representação matemática da regressão linear e nela pode-se identificar as variáveis y, a, x e b. O coeficiente ’y’ representa o termo independente que deseja-se classificar ou prever, ’x’ a variável dependente, ’a’ a inclinação da reta no plano cartesiano e b a constante que define o valor de ’y’ quando ’x’ é 0.

5.2 ALGORÍTIMO DE REGRESSÃO MULTIVARIÁVEL

Modelos de regressão linear são abordagens simples do aprendizado supervisionado. Neles, a relação entre as variáveis dependentes e independentes pode ser expressado através de uma reta, ou algo bem próximo a isso (MENON, 2017b).

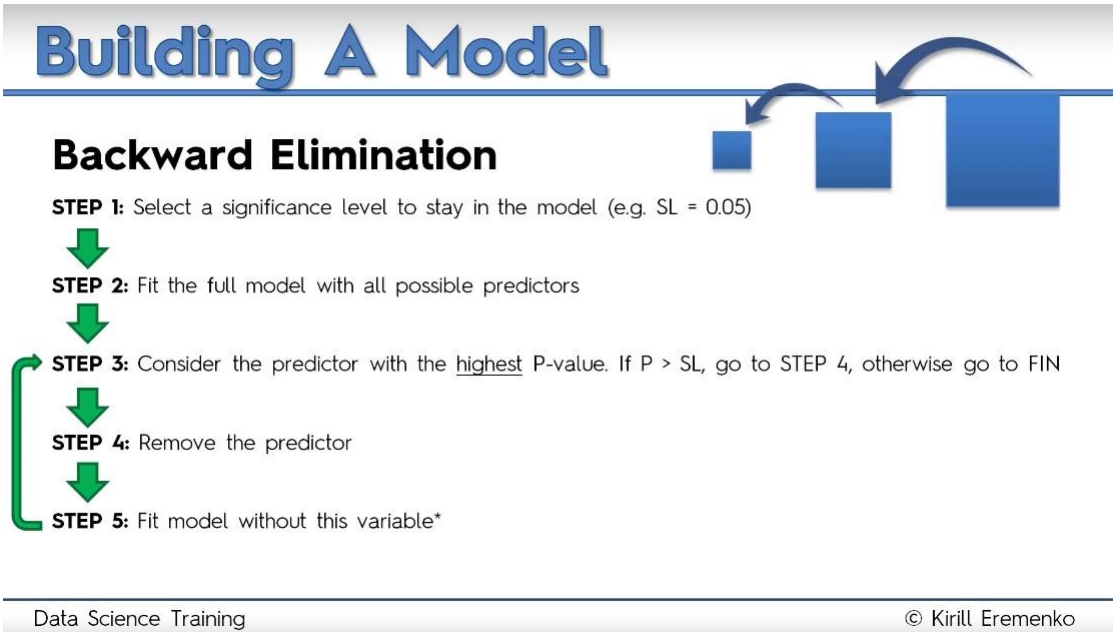
O algorítimo de regressão linear simples propõe um contexto onde é possível criar previsões através da análise de apenas uma variável dependente e uma independente, que pode ser expressada através da função y = f(x). Porém em grande parte dos problemas a variável dependente necessita de mais de uma variável independente para ser expressada, estes são os chamados modelos de regressão linear multivariável (MENON, 2017b).

Com o intuito de obter o melhor modelo possível foi utilizada a técnica de *Backwards Elimination* para elaboração do modelo de regressão linear multivariável. Para compreender este mecanismo é necessário entender o conceito de *p-value*, que é uma medida estatística que auxilia o cientista a avaliar a exatidão de suas respostas.

O *p-value* é utilizado para determinar se os resultados do experimento realizados estão dentro de um mesmo alcance de valores para o evento observado. O cientista determina um *p-value* ideal antes da realização de seus testes, e comumente ele é determinado em 0,05. Caso o valor *p-value* analisado esteja abaixo de 0,05, então a hipótese nula é descartada, em outras palavras a hipótese de que aquela variável não possui um efeito significativo nos resultados é rejeitada (HOW, 2019).

O algorítimo de *Backwards Elimination* representada na figura 9 demonstra os passos a serem seguidos para retirar as variáveis do modelo que não possuem valor significativo no resultado do experimento.

Figura 9 – Algorítimo de *Backwards Elimination*



Fonte: Disponível em: <https://www.kaggle.com/srisudheera/backward-elimination>. Acesso em: 22 de maio de 2019.

Passo 1: Definir o *p-value*. Para nosso experimento foi definido o *p-value* em 0,05. Passo 2: Modelar o problema com todas as variáveis disponíveis. Passo 3: Após aplicar o algorítimo de *machine learning*, identificar a variável independente com o *p-value* mais alto e ir para o passo 4. Caso não haja preditores com *p-value* maiores do que 0,05 vá para o passo 6. Passo 4: Remover a variável independente com o *p-value* mais elevador. Passo 5: Modelar o problema com todas as variáveis independentes restantes e voltar para o passo 3. Passo 6: Fim.

5.3 SUPPORT VECTOR REGRESSION (SVR)

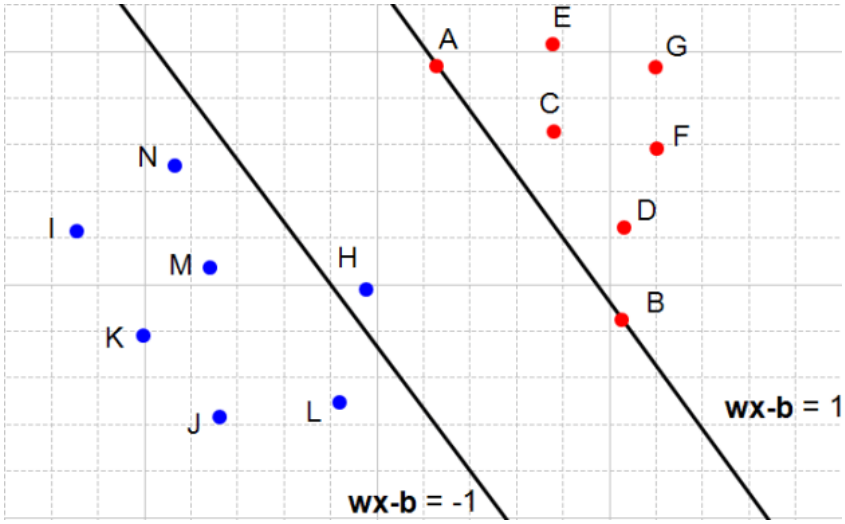
Support Vector Machines (SVM) são máquinas de aprendizado que implementam o princípio indutivo de minimização do risco estrutural para obter boas generalizações em um limitado número de padrões de aprendizado. Minimização do risco estrutural (SRM) envolve tentativas simultâneas de minimizar o risco empírico e a dimensão (BASAK; PAL; PATRANABIS, 2007). Originalmente a teoria de SVMs foi desenvolvida por *Vapnik* e colegas nos laboratórios da *ATT Bell* (VAPNIK; SMOLA; GOLOWICH, 1997). O algoritmo SVM consegue reconhecer padrões sutis em data sets complexos, performando um aprendizado de classificação discriminativa por exemplos para então prever dados que nunca foram apresentados ao algoritmo, por este motivo pode-se classificá-lo como um algoritmo de aprendizado supervisionado. Performance consideráveis foram obtidas em aplicação de predições de regressão e time series (DRUCKER, 1997). Uma SVM pode ser utilizada tanto para problemas de regressão

quanto para problemas de classificação. As próximas sub sessões explicam o funcionamento básico de uma SVM e uma breve fundamentação matemática sobre SVR que é a extensão de uma SVM para problemas de regressão.

5.3.1 Funcionamento Básico

Uma SVM constrói um hiperplano ou um conjunto de hiperplanos em um espaço dimensional grande ou infinito. Uma boa separação é alcançada com o hiperplano que possui a maior distância para o ponto de treinamento mais perto de qualquer classe, considerando que no geral, quanto maior a margem menor será a generalização do erro do classificador. Na figura 10 pode-se notar o hiperplano e a margem. (DEVELOPERS (B), 2019)

Figura 10 – Representação do SVM no hiperplano



Fonte: Disponível em: <http://www.decom.ufop.br/imobilis/svm-entendendo-sua-matematica-parte-3-o-hiperplano-otimo/>. Acesso em: 22 de maio de 2019.

5.3.2 Formulação Matemática

Dado o vetor de treinamento Xi n, i =1,....,n, e o vetor y n, a SVR resolve o problema primal representado na figura 11, e dual representado na figura 12, onde "e"é o vetor de todos, C > 0 é a barreira superior, Q é uma matriz n por n semi definida positiva,Qij K(xi,xj/) = (xi)(xi)T(xj) é o *kernel*, e o vetor de treinamento encontra-se implicitamente mapeado em um espaço de dimensão grande pela função. (DRUCKER, 1997). A função de decisão é representada na figura 13

Figura 11 – Problema primal

$$\begin{aligned} \min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \\ \text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i, \\ w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n \end{aligned}$$

Fonte: Feita pelo autor

Figura 12 – Problema dual

$$\begin{aligned} \min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon e^T (\alpha + \alpha^*) - y^T (\alpha - \alpha^*) \\ \text{subject to } e^T (\alpha - \alpha^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, n \end{aligned}$$

Fonte: Feita pelo autor

Figura 13 – Função de decisão

$$\begin{aligned} \min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon e^T (\alpha + \alpha^*) - y^T (\alpha - \alpha^*) \\ \text{subject to } e^T (\alpha - \alpha^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, n \end{aligned}$$

Fonte: Feita pelo autor

5.4 *RANDOM FOREST REGRESSION* (RFR)

Random Forest é um método de *machine learning* supervisionado que pode ser usado tanto para regressão quanto para classificação, (BREIMAN, 2001) propôs random forest pela primeira vez. Random forest adiciona uma camada de aleatoriedade ao método bagging descrito por (BREIMAN, 1996). Além de construir cada árvore usando uma amostra de *bootstrap* diferente dos dados, as *Random Forest* alteram como as árvores de regressão e classificação são construídas. Em uma *Random Forest* padrão, cada nó é separado usando a melhor separação ao longo de todas as

variáveis. Na *Random Forest* cada nó é separado usando a melhor separação ao longo dos subconjunto de preditores aleatoriamente escolhidos a partir do nó em questão. Essa estratégia contra intuitiva possui uma performance boa se comparada a outros classificadores, e ela é robusta contra *overfitting*. (LIAW; WIENER, 2002)

5.4.1 Algorítimo

O algoritmo de *Random Forest* para classificação e regressão segue os seguintes passos (LIAW; WIENER, 2002)

- Desenhe n árvores com amostras a partir dos dados originais.

- Para cada amostra, cresça uma árvore de classificação ou regressão sem podas, com a seguinte modificação de uma árvore de regressão/classificação padrão: para cada nó, ao invés de escolher a melhor forma ao longo de todos preditores, aleatoriamente pega-se uma amostra *mtry* dos preditores e escolha a melhor divisão dentre as variáveis. Realize a predição do novo dado agregando as predições das n árvores (Votos majoritários para classificação ou média para regressão)

Uma estimativa da taxa de erro pode ser obtida baseada nos dados de treino com o seguinte:

- A cada iteração do *bootstrap*, faça a predição do dado que não está na amostra *bootstrap*, usando a árvore que cresceu como amostra *bootstrap*.

- Agregue a predição que foi feita com os dados fora da amostra *bootsrap*, os chamados OOB ou dados *"out-of-bag"*. Calcule a taxa de erro, e chame de estimativa OOB de taxa de erro (LIAW; WIENER, 2002).

6 METODOLOGIA DE PESQUISA

Neste capítulo é discutida a metodologia de pesquisa. Conforme discutido anteriormente, a maior dificuldade dos investidores são seus próprios instintos e emoções, que visam os ganham a curto prazo, fato que cega parte de suas decisões de investimento e impede a criação de estratégias eficazes que sejam capazes de retornar lucros.

Para possibilitar o entendimento do leitor sobre o objetivo deste trabalho de conclusão de curso e qual problema ele busca resolver, foram apresentados nos capítulos posteriores conceitos sobre o surgimento do mercado financeiro, análise técnica, indicadores técnicos e ML.

6.1 FERRAMENTAS UTILIZADAS

Nesta seção são apresentadas todas as ferramentas utilizadas para o desenvolvimento deste trabalho de conclusão de curso, tanto da parte de programação quanto da parte de escrita.

6.1.1 Python

A linguagem de programação utilizada para este trabalho foi o *Python*. Ele foi criado em 1989 por *Guido van Rossum* no Instituto de Pesquisa Nacional para Matemática e Ciência da Computação nos países baixos. (YEGULALP, 2018)

O *Python* é uma linguagem de programação de alto nível, orientada a objeto e de script. Ao lado do R é uma das linguagens mais utilizadas para ciência de dados e possui diversas bibliotecas que auxiliam os cientistas em diversas aplicações. (YEGULALP, 2018)

O *Python* pode ser utilizada para propósitos gerais, e possui tipagem dinâmica, e uma das suas principais características é a sua fácil legibilidade e pouca verbosidade, ou seja seus códigos são menores se comparadas a outras linguagens de alto nível como C ou Java. (YEGULALP, 2018)

6.1.2 Scikit-learn

O *Scikit-learn* também conhecido como sklearn é uma biblioteca escrita ma sua maior parte em python e com alguns algorítimos escritos em *Cython* com o intuito

de otimizar o desempenho. David Cournapeau foi seu idealizador em um projeto do *Google Summer of Code*. (DEVELOPERS (B), 2019)

O *Scikit-learn* reúne diversos modelos de *machine learning* já implementados através de diversas classes e métodos. Sua documentação e guias de uso podem ser encontrados em (DEVELOPERS (B), 2019).

6.1.3 Pandas

Pandas é uma biblioteca *open-source* de fácil uso que provém alta performance para análise de estrutura de dados. (AGUIAR, 2018)

É focada para programadores *Python* e fornece diversas ferramentas que permitem manipular os data sets. Suas classes provém utilidades para o pré-processamento de dados que é a primeira fase de aplicação de diversos modelos de *macinhe learning*. (AGUIAR, 2018)

6.1.4 Anaconda

Anaconda é a plataforma de data *science* mais popular do mundo com mais de 13 milhões de usuários. O Anaconda simplifica e automatiza a colaboração e implementação de *ata science* e ML com alta performance e de forma escalável. (CUSTODIO, 2017)

6.1.5 Spyder

Spyder é uma das diversas *Integral Development Enviroment* (IDE) presentes na plataforma Anaconda. IDEs são utilizadas para fazer a tradução da linguagem utilizada pelo programador para o baixo nível, para possibilitar que o computador entenda a lógica descrita. Ele é uma poderosa ferramenta desenvolvida em *python* por cientista de dados para que outros cientistas possam desenvolver suas aplicações. (CUSTODIO, 2017)

6.1.6 Overleaf

Overleaf é uma ferramenta online de escrita que facilita a redição de textos científicos, e foi utilizada para a confecção deste trabalho. (BASU, 2016)

7 DESENVOLVIMENTO

Nesta sessão são discutidos os passos seguidos para a realização do experimento que objetiva prever o preço de ativos do mercado financeiro através do uso de parâmetros técnicos selecionados por (HEGAZY, 2013) em seu trabalho.

7.1 DATASET

Datasets são coleções de itens relacionados que podem ser acessados de forma individual ou em combinação como o todo de uma entidade. Os *datasets* são organizados através de estrutura de dados (SHARMA, 2019).

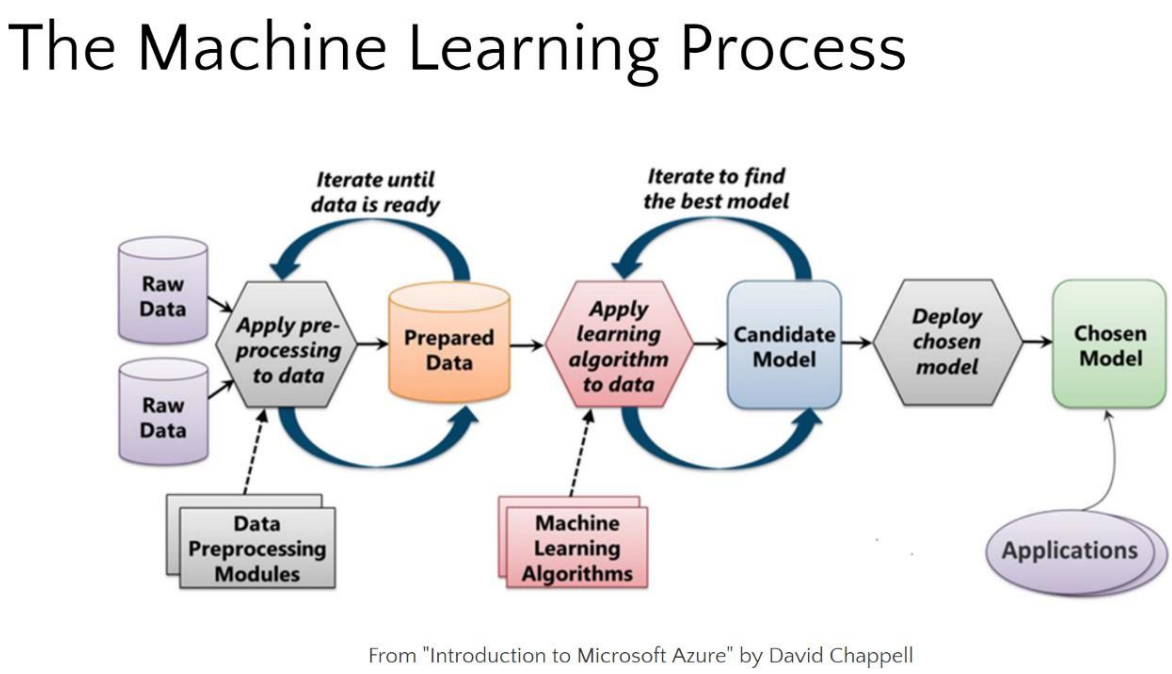
Os bancos de dados podem ser considerados um exemplo de *dataset*, onde cada coluna é o item de uma determinada entidade, e a coleção dos itens descreve diversas características atribuídas a ela (SHARMA, 2019).

O primeiro passo para a realização do experimento foi a coleta de dados no site https://finance.yahoo.com/. Nele é possível encontrar uma grande variedade de informações estatísticas e gráficas da análise técnica. Os indicadores técnicos escolhidos para este experimento foram baseados no estudo de (HEGAZY, 2013), sendo eles o *Money Flow Index*, *Relative Strength Index*, *Stochastic Oscilator* e *Moving Average Convergence Divergenc*.Foram colhidos dados do período de 2012-2019 da distribuidora Ambev na bolsa de valores brasileira Ibovespa. Os dados de cada ativo foram salvos em um arquivos .csv. Os preços são os valores de cada fechamento mensal da ação.

7.2 DATA PREPROCESSING

O *Data PreProcessing* é uma técnica muito utilizada na mineração de dados, que objetiva transformar dados brutos em um formato estruturado. No mundo ideal todos os dados estão completos, consistentes, e possuem tendências claras com poucos erros. Porém devida a grande divergência entre o idealismo e a realidade, surge a necessidade do tratamento dos dados, para aproximar seu aspecto bruto do ideal. A Figura 14 representa processo de ML, desde o *Data PreProcessing* até a escolha final do modelo (SHARMA, 2019).

Figura 14 – Representação do processo de *machine learning*



Fonte: Disponível em: <https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa>. Acesso em: 21 de abril de 2019.

Para tratar os dados colhidos foi utilizada a técnica de *Data Preprocessing* descritos a seguir: (SHARMA, 2019)

- Importar as bibliotecas *matplotlib* e pandas para manipulação dos dados (SHARMA, 2019).

- Importar o *dataset* através da biblioteca Pandas.

- Verificar dentro do *dataset* se existem valores ausentes.

- Verificar dentro do *dataset* os valores categóricos. Dentro de uma massa de dados, campos que possuem características como nome, cor ou sexo devem ser tratados e transformados em números para possibilitar a modelagem do problema.

- Fazer o *Feature Scaling*, que significa colocar os dados do *dataset* na mesma escala, para que uma variável não sobreponha a outra dentro da aplicação do modelo.

- Dividir o *dataset* em um conjunto de treino e o conjunto de teste.

7.3 ALGORÍTIMOS

7.3.1 Aplicar o algorítimo ao modelo candidato

Esse passo consiste na escolha dos algorítimos de aprendizado utilizados no experimento. A biblioteca *Scikit-learn* ou sklearn, disponibiliza inúmeras classes e métodos que implementam diversos algorítimos de ML, como a regressão linear multivariável, *support vector regressor, support vector machine, decision tree regression* entre outros.

Com as variáveis dependentes e independentes do *dataset* divididos em um conjunto de treinamento e teste, é necessário apenas instanciar as classes que implementam os algorítimos e utilizar o método *fit* para preencher o modelo com os dados de entrada e *predict* para gerar a saída com os resultados da previsão.

Para a implementação da RLM, foi utilizada o algorítimo de *Backwards Elimination* descrito no capítulo 5 para verificar a existência de variáveis independentes sem relevância para os resultados do experimento.

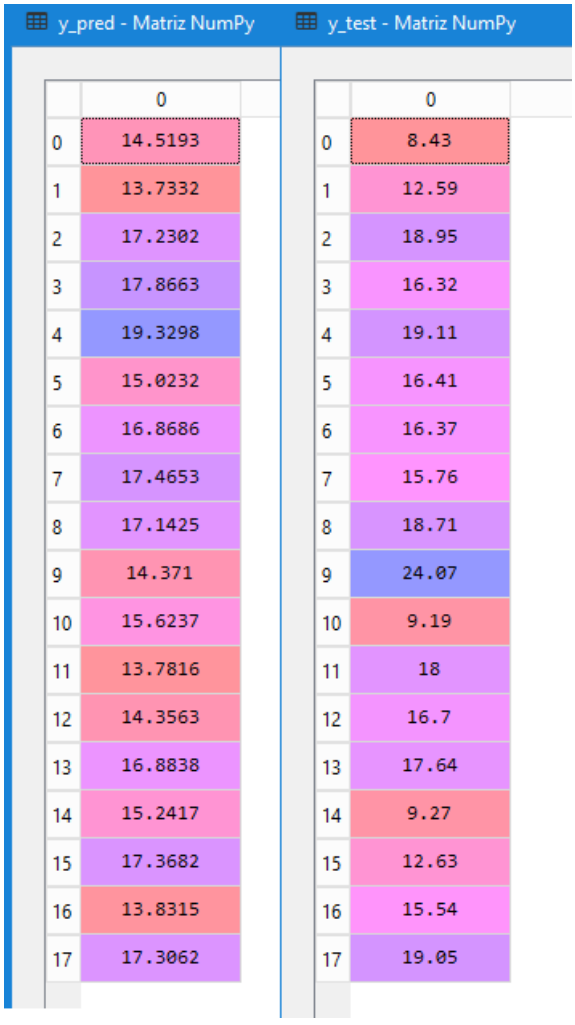
Após implementado o primeiro algorítimo, obteve-se um *teamplate* genérico que foi utilizado para a implementação dos demais experimentos. Os resultados são discutidos no capítulo 8.

8 RESULTADOS

8.1 REGRESSÃO LINEAR MÚLTIPLA APLICADA À AMBEV

A tabela da esquerda representada na figura 15 mostra os valores de saída da variável independente ypred calculados pela regressão, enquanto a tabela da direita demonstra o ytest, que foi dividido em treinamento e teste na fase de *Data Preprocessing*. O *dataset* foi dividido em 80 por cento treinamento e 20 por cento teste

Figura 15 – Resultado da regressão vs Valor real



	0
0	14.5193
1	13.7332
2	17.2302
3	17.8663
4	19.3298
5	15.0232
6	16.8686
7	17.4653
8	17.1425
9	14.371
10	15.6237
11	13.7816
12	14.3563
13	16.8838
14	15.2417
15	17.3682
16	13.8315
17	17.3062

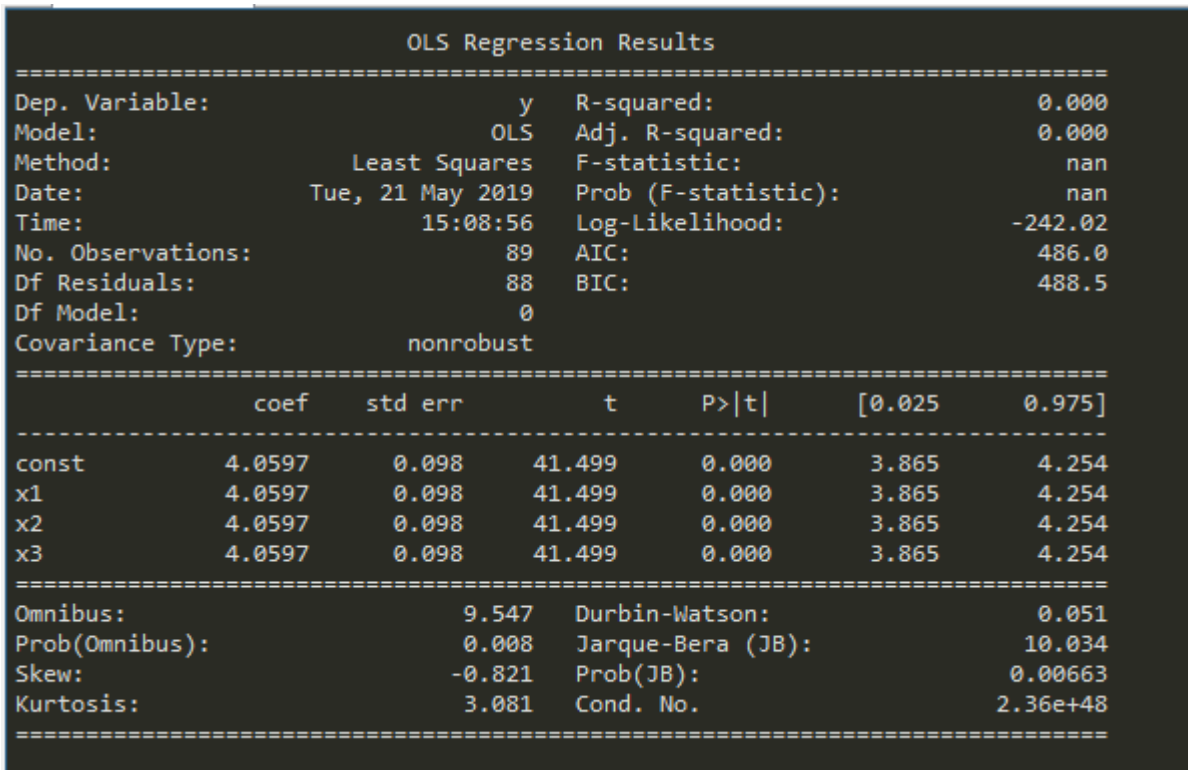
	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

Fonte: Feita pelo autor

Ao analisar os resultados obtidos, pode-se notar que o algorítimo foi pouco assertivo na maior parte dos casos, apesar do fato de que, nas linhas 4 e 6 ele aproximou do valor real.

Para tentar melhorar o modelo, foi aplicado A técnica de *Backward Elimination* para identificar as variáveis independentes com *p-value* maiores do que 0,05. O resultado esta representado na figura 16.

Figura 16 – Resultado após aplicação do *Backward Elimination*



OLS Regression Results

Dep. Variable:	y	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	nan
Date:	Tue, 21 May 2019	Prob (F-statistic):	nan
Time:	15:08:56	Log-Likelihood:	-242.02
No. Observations:	89	AIC:	486.0
Df Residuals:	88	BIC:	488.5
Df Model:	0		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0597	0.098	41.499	0.000	3.865	4.254
x1	4.0597	0.098	41.499	0.000	3.865	4.254
x2	4.0597	0.098	41.499	0.000	3.865	4.254
x3	4.0597	0.098	41.499	0.000	3.865	4.254

Omnibus:	9.547	Durbin-Watson:	0.051
Prob(Omnibus):	0.008	Jarque-Bera (JB):	10.034
Skew:	-0.821	Prob(JB):	0.00663
Kurtosis:	3.081	Cond. No.	2.36e+48

Fonte: Feita pelo autor

Como pode-se ver, o p-value representado pela coluna P>|t| de todos os coeficientes independentes é igual a 0, portanto seguindo a técnica de *BackWard Elimination* nenhuma variável deve ser retirada do modelo. Com os resultados obtidos neste experimento, pode-se concluir que o uso do algorítimo de RLS aplicado a este modelo é pouco assertivo na predição de valores do mercado financeiro.

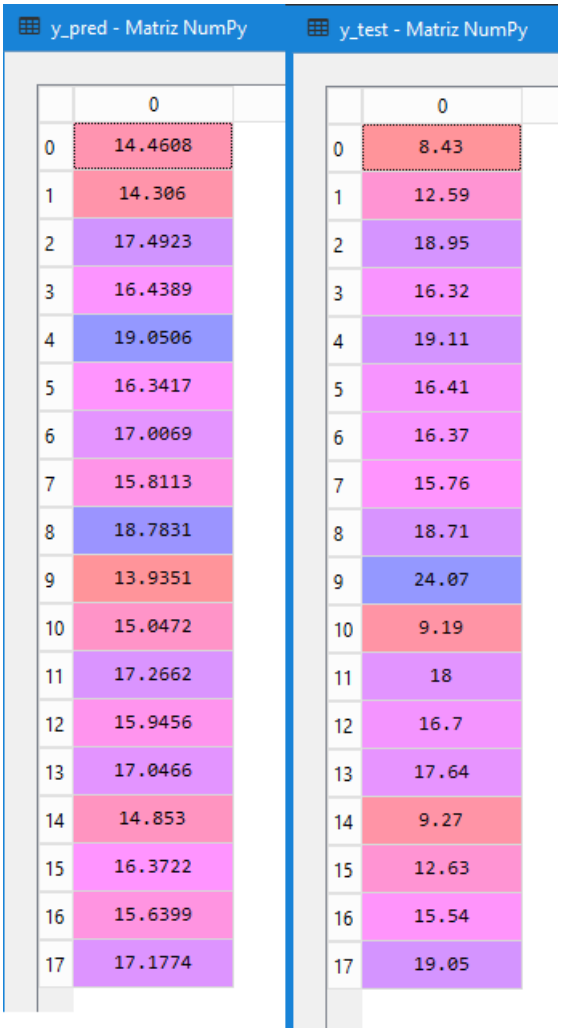
Devido ao fato do mercado financeiro possuir um comportamento extremamente imprevisível e complexo, algorítimos que dependem de linearidade são pouco assertivos para a sua predição.

8.2 SVR APLICADO À AMBEV

A tabela da esquerda representada na figura 17 mostra os valores de saída da variável independente ypred calculados pela regressão, enquanto a tabela da

direita demonstra o ytest, que foi dividido em treinamento e teste na fase de Data Preprocessing. O *dataset* foi dividido em 80 por cento treinamento e 20 por cento teste.

Figura 17 – Resultado da regressão vs Valor real SVM



y_pred - Matriz NumPy

	0
0	14.4608
1	14.306
2	17.4923
3	16.4389
4	19.0506
5	16.3417
6	17.0069
7	15.8113
8	18.7831
9	13.9351
10	15.0472
11	17.2662
12	15.9456
13	17.0466
14	14.853
15	16.3722
16	15.6399
17	17.1774

y_test - Matriz NumPy

	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

Fonte: Feita pelo autor

Ao analisar os resultados e comparando com os valores obtidos no experimento anterior, foi constatado que a assertividade do SVR teve um resultado consideravelmente melhor.

Quando a variável independente assume valores entre 15 à 19, a assertividade do modelo é consideravelmente maior que nos demais. Periodicamente o mercado financeiro tem altas e baixas oriundas de otimismo infundado ou pessimismo inexplicável. Tal fato, causa variações no preço do ativo que não puderam ser identificadas através da aplicação SVR utilizando os parâmetros técnicos selecionados.

Casos em que a variável dependente assumiu valores como 9,27 e 24, tiveram suas predições altamente incorretas, pois são valores anômalos aos demais e as causas destas anomalias não puderam ser identificadas através do aplicação do SVR aos parâmetros técnicos.

O SVR possui uma boa assertividade em mercados de consolidação onde o ativo não sofre altas ou quedas bruscas de um período ao outro.

8.3 RFR APLICADO À AMBEV

A tabela da esquerda representada na figura 18 mostra os valores de saída da variável independente ypred calculados pela regressão, enquanto a tabela da direita demonstra o ytest, que foi dividido em treinamento e teste na fase de *Data Preprocessing*.

Figura 18 – Resultado da regressão vs Valor real RFR

y_pred - Matriz NumPy

	0
0	11.43
1	12.333
2	17.809
3	16.13
4	19.101
5	16.945
6	15.481
7	16.264
8	19.042
9	11.053
10	13.291
11	15.574
12	16.154
13	16.622
14	14.914
15	15.583
16	16.064
17	17.814

y_test - Matriz NumPy

	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

Fonte: Feita pelo autor

Ao analisar os resultados obtidos neste experimento, pode-se notar uma assertividade superior ao da regressão linear multivariável, assim como na implementação do SVR.

Comparando os resultados obtidos do experimento utilizando o algorítimo Random Forest Regression (RFR) com o SVR, pode-se notar um nível de assertividade similar. Assim como no SVR, houve falha na predição de valores muito altos ou muito baixos, como representados na linha 9, 10 e 14.

Assim como o SVR o RFR teve um bom desempenho na maior parte da predição, porém devido as anomalias do mercado que podem ser oriundas do pessimismo inexplicável ou otimismo infundado, grandes oscilações do preço não tiveram suas previsões corretas. Os parâmetros técnicos selecionados para criação do modelo não foram suficientes para detectar tais anomalias.

Pode-se concluir que o algorítimo RFR aplicado ao modelo escolhido possui um bom desempenho em um cenários de consolidação da ação, onde o preço se mantém estável, porém possui baixa assertividade em mercados de clara tendência de alta ou baixa.

9 CONCLUSÃO

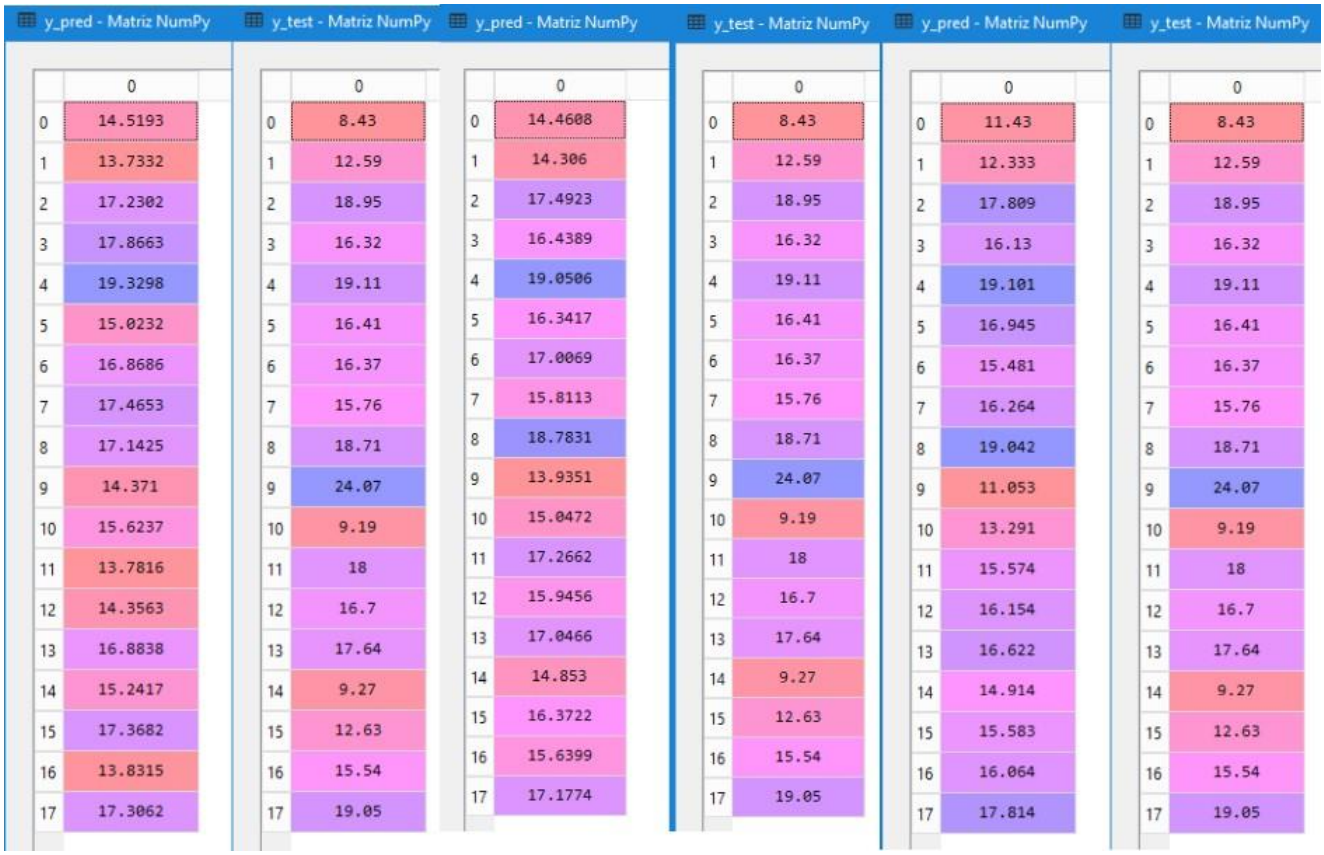
O mercado financeiro se tornou tão popular devido a possibilidade de obtenção de lucro através da flutuação de preço dos ativos no curto prazo. O valor de uma ação pode decolar ou despencar em questão de meses e como consequência o dinheiro dos investidores pode se multiplicar ou dividir no mesmo período. Para evitar perdas é necessário buscar conhecimento e entender as forças que regem o comportamento do mercado financeiro.

O trabalho aqui realizado buscou apresentar bases empíricas de diversas áreas que compõe o mercado financeiro, citando fatos históricos que ocasionaram seu surgimento, os estudos da análise técnica que através de dados estatísticos tentam prever o comportamento do valor das ações, a economia comportamental, que busca prever o comportamento em massa dos investidores que compõe o mercado e a implementação de algorítimos de regressão linear e não linear aos parâmetros técnicos.

Ao analisar os resultados da aplicação do algorítimo de regressão linear múltipla aos parâmetros técnicos escolhidos, pudemos enxergar uma assertividade baixa na predição, devido ao comportamento complexo do mercado que dificilmente pode ser descrito através de uma abordagem linear.

Já os algorítimos SVR e RFR obtiveram resultados mais próximos dos reais, porém ambos falharam na predição de valores extremos da amostra de dados. Através da análise dos resultados obtidos, pode-se concluir que o investidor precisa entender a tendência do mercado e quais forças estão atuando nele naquele momento para possibilitar a escolha da melhor abordagem que objetiva prever o seu comportamento.

Figura 19 – ML vs SVR vs RFR



	0
0	14.5193
1	13.7332
2	17.2302
3	17.8663
4	19.3298
5	15.0232
6	16.8686
7	17.4653
8	17.1425
9	14.371
10	15.6237
11	13.7816
12	14.3563
13	16.8838
14	15.2417
15	17.3682
16	13.8315
17	17.3062

	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

	0
0	14.4608
1	14.306
2	17.4923
3	16.4389
4	19.0506
5	16.3417
6	17.0069
7	15.8113
8	18.7831
9	13.9351
10	15.0472
11	17.2662
12	15.9456
13	17.0466
14	14.853
15	16.3722
16	15.6399
17	17.1774

	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

	0
0	11.43
1	12.333
2	17.809
3	16.13
4	19.101
5	16.945
6	15.481
7	16.264
8	19.042
9	11.053
10	13.291
11	15.574
12	16.154
13	16.622
14	14.914
15	15.583
16	16.064
17	17.814

	0
0	8.43
1	12.59
2	18.95
3	16.32
4	19.11
5	16.41
6	16.37
7	15.76
8	18.71
9	24.07
10	9.19
11	18
12	16.7
13	17.64
14	9.27
15	12.63
16	15.54
17	19.05

Fonte: Feita pelo autor

Os parâmetros técnicos escolhidos para compor os modelo de estudo não foram eficientes para prever valores da extremidade mais alta e baixa da amostra de dados, porém tiveram um bom desempenho dentro dos valores médios. Os algorítimos de SVR e RFR tiveram boa assertividade em momentos de consolidação da ação, onde o preço do ativo se manteve estável.

Os estudos e análises aqui apresentados não devem ser utilizados isoladamente para a criação de estratégias de operação em mercado financeiro. É necessário a aprofundar do conhecimento de cada tópico apresentado neste trabalho e entender as relações entre as forças que influenciam o mercado financeiro para ser capaz de criar modelos que possam prever de forma eficaz o seu comportamento.

REFERÊNCIAS

AGUIAR, V. Disponível em: <https://medium.com/data-hackers/uma-introdumaio 2018>. Acesso em: 14 de junho 2019.

BASAK, D.; PAL, S.; PATRANABIS, D. C. Support vector regression. v. 11, outubro. 2007.

BASU, A. How to use overleaf to write your papers: Part i: Basic minimalist setup. julho 2016. Disponível em: <https://medium.com/thoughts-philosophy-writing/ how-to-use-overleaf-to-write-your-papers-part-i-basic-minimalist-setup-6599268c095f>. Acesso em: 14 de junho 2019.

BREIMAN, L. Bagging predictors. v. 45, p. 123–140, outubro 1996.

BREIMAN, L. Random forests. v. 45, p. 5–32, outubro 2001.

CASTRO, M. As 21 melhores frases de warren buffett sobre investimentos. fev. 2019. Disponível em: <https://www.infomoney.com.br/carreira/gestao-e-lideranca/noticia/ 7923063/as-21-melhores-frases-de-warren-buffett-sobre-investimentos>. Acesso em: 17 de junho 2019.

CIALDINI, R. B. **As Armas da Persuasão: Como Influenciar e Não Se Deixar Influenciar**. [S.l.]: GMT Editores Ltda., 2009. v. 3. 25-50 p.

CUSTODIO, S. Python, anaconda e spyder. fev. 2017. Disponível em: <https: //en.wikipedia.org/wiki/Anaconda_(Python_distribution)>. Acesso em: 14 de junho 2019.

DEVELOPERS (A), S. L. 1.4. support vector machines. Cambridge, 2019. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: 30/05/2019.

DEVELOPERS (B) scikit learn. Documentation of scikit-learn 0.21.2. junho 2019. Disponível em: <https://scikit-learn.org/stable/documentation.html>. Acesso em: 14 de junho 2019.

DRUCKER, H. Support vector regression machines. Cambridge, p. 155–161, 1997.

HEGAZY, O. Technical analysis of stocks and trends definition. EUA, dez. 2013. Disponível em: <https://www.researchgate.net/publication/259240183_A_Machine_ Learning_Model_for_Stock_Market_Prediction>. Acesso em: 05 abril 2019.

HOW, W. How to calculate p value. EUA, mar. 2019. Disponível em: <https: //www.wikihow.com/Calculate-P-Value>. Acesso em: 22 de maio 2019.

JAMES, C. (A) Relative strength index - rsi definition. EUA, fev. 2019. Disponível em: <https://www.investopedia.com/terms/r/rsi.asp>. Acesso em: 05 de abril 2019.

JAMES, C. Technical analysis of stocks and trends definition. EUA, fev. 2019. Disponível em: <https://www.investopedia.com/terms/t/technical-analysis-of-stocks-and-trends. asp>. Acesso em: 19 de março 2019.

KENTON, W. Behavioral economics. EUA, set. 2017. Disponível em: <https: //www.investopedia.com/terms/b/behavioraleconomics.asp>. Acesso em: 19 de março 2019.

KENTON, W. Financial market definition. EUA, mar. 2019. Disponível em: <https://www.investopedia.com/terms/f/financial-market.asp>. Acesso em: 08 de abril 2019.

KUEPPER, J. Technical analysis: Indicators and oscillators. abr. 2017. Disponível em: <https://www.investopedia.com/university/technical/techanalysis10.asp>. Acesso em: 29 de março 2019.

LIAW, A.; WIENER, M. Classification and regression by randomforest. outubro 2002.

LORENA, A. d. C. A. C. Uma introdução às support vector machines. 2007. Disponível em: <https://seer.ufrgs.br/rita/article/viewFile/rita_v14_n2_p43-67/3543>. Acesso em: 21 abril 2019.

MENON, P. Data science simplified part 4: Simple linear regression mo- dels. EUA, agos 2017. Disponível em: <https://towardsdatascience.com/ data-science-simplified-simple-linear-regression-models-3a97811a6a3d>. Acesso em: 28 de abril 2019.

MENON, P. Data science simplified part 5: Multivariate regression mo- dels. EUA, agos 2017. Disponível em: <https://towardsdatascience.com/ data-science-simplified-part-5-multivariate-regression-models-7684b0489015>. Acesso em: 28 de abril 2019.

MITCHELL, C. Money flow index - mfi definition and uses. EUA, abril 2019. Disponível em: <https://www.investopedia.com/terms/m/mfi.asp>. Acesso em: 18 de maio 2019.

RADAR, T. Uma cartilha sobre o macd. 2018. Disponível em: <https://www.tororadar. com.br/investimento/analise-tecnica/macd>. Acesso em: 26 mar. 2019.

SHARMA, M. What steps should one take while doing data prepro- cessing? EUA, fev. 2019. Disponível em: <https://hackernoon.com/ what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa>. Acesso em: 22 de maio 2019.

THALER, R. **Misbehaving**. [S.l.: s.n.], 2011. v. 3. 20-80 p. Acesso em: 05 abril 2018.

VAPNIK, V.; SMOLA, A.; GOLOWICH, S. E. Support vector method for function approximation, regression estimation, and signal processing. Cambridge, v. 9, outubro. 1997.

VERSIGNASSI, A. **Crash**. [S.l.]: GMT Editores Ltda., 2009. v. 3. 20-80 p.

YEGULALP, S. Python. junho 2018. Disponível em: <https://www.infoworld.com/article/ 3204016/what-is-python.html>. Acesso em: 14 de junho 2019.