INSTITUTO FEDERAL DE SANTA CATARINA

KLEITON CARLOS DE SOUZA

Predição de variação de preços de criptomoedas utilizando Aprendizado de Máquina

São José - SC

Novembro/2018

PREDIÇÃO DE VARIAÇÃO DE PREÇOS DE CRIPTOMOEDAS UTILIZANDO APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso apresentado à Coordenadoria do Curso de Engenharia de Telecomunicações do campus São José do Instituto Federal de Santa Catarina para a obtenção do diploma de Engenheiro de Telecomunicações.

Orientador: Ramon Mayor Martins

São José - SC

Novembro/2018

RESUMO

Este trabalho procura mostrar quais os melhores métodos de análise de dados e técnicas de aprendizado de máquina para realizar a predição de variação de preços de criptomoedas. Com o propósito de fazer uma análise do histórico de dados de algumas criptomoedas e organizar os dados gerando uma predição de valores futuros, podendo então eleger uma criptomoeda que esteja com uma tendência de alta (ou baixa), mostrando assim qual delas pode ser mais rentável em determinado período.

Palavras-chave: Criptomoedas. Predição. Aprendizado de máquina. Cryptocurrency.

ABSTRACT

This paper aims to show which the best methods of data analysis and machine learning techniques are most effective in predicting the price variation of cryptomoedas. For the purpose of analyzing the data history of some cryptocoins, and with reliability and stability, organize the data generating a prediction of future values, being able to choose a cryptocoins with a trend that is in a high (or low), thus showing which may be more profitable in a given period.

Keywords: Cryptocoins. Prediction. Machine Learning. Cryptocurrency.

SUMÁRIO

1 INTRODUÇÃO 7
1.1 Objetivos 7
1.1.1 Objetivo Geral 7
1.1.2 Objetivo Específico 7
1.2 Organização do Texto 8

2 FUNDAMENTAÇÃO TEÓRICA 9
2.1 Criptomoedas 9
2.1.1 Bitcoin 9
2.1.2 Ethereum 9
2.1.3 Ethereum Classic 9
2.1.4 Monero 10
2.1.5 Litecoin 10
2.2 Técnicas Aprendizado de Máquina 10
2.2.1 Regressão Linear Simples 10
2.2.2 Regressão Logística 11
2.2.3 Regressão Bayesiana 11
2.2.4 Random Forest 12

3 METODOLOGIA 13
3.1 Base de Dados 13
3.2 Aparato Técnico 14
3.2.1 Linguagem R 14

4 PROPOSTA 15
4.1 Cronograma 15

REFERÊNCIAS 17

1 INTRODUÇÃO

Com a crescente das criptomoedas nos últimos anos, sua mineração tem tomado força e se mostrado um bom investimento de longo prazo, para quem dispõe do hardware necessário para tal. A mineração é feita por programadores que fornecem poder computacional para encontrar a chave que criptografa os blocos e também fazem o registro de suas transações. Essa chave é chamada de *hash* e a tarefa de encontrá-la é feita através da resolução de cálculos criptográficos que validam as transações. Sempre que uma chave correta é encontrada, o minerador recebe uma recompensa pelo trabalho. Sendo assim, subentende-se que quanto maior o seu poder computacional, maior a chance de você encontrar uma chave (CARLOS, 2018).

O alto investimento em equipamentos para mineração de criptomoedas (chamados *rigs* de mineração), é devido ao alto poder computacional que é necessário para quebrar as chaves criptográficas das moedas encontradas. Então, para que seja rentável, uma predição da flutuação do valor das criptomoedas torna-se atrativa para quem quer investir. As criptomoedas possuem valor flutuante, não há interferência de agências financeiras e não está atrelada a fatores econômicos de determinado país, fazendo com que não sofra tanta influência por conta de crises econômicas, políticas ou cotações de moedas tradicionais. Sendo influenciada única e exclusivamente por ações dos vendedores e compradores, tem relação direta com oferta e demanda (SCHIAVON, 2017).

Como a variação dos valores das criptomoedas não está atrelada a um indicador expressivo, podemos aplicar métodos de Big Data (regressão e análise de dados) e técnicas de Inteligência Artificial (aprendizado de máquina), dois dos pilares que impulsionam o desenvolvimento de tecnologias inovadoras (BARBOSA, 2017). Aplicando esses métodos às informações contidas nos *datasets* para apresentar uma tendência da variação do valor das criptomoedas. As técnicas de aprendizado de máquina (do inglês, *Machine Learning*), são caracterizadas por investigar como as máquinas podem adquirir conhecimento através da extração de padrões a partir de um conjunto de dados (RUSSEL; NORVIG, 2013), buscando o desenvolvimento de algoritmos que permitam que as máquinas possam se tornar capazes de tomar decisões com autonomia. Podemos então gerir um comitê para delegar ou eleger qual criptomoeda está com uma tendência melhor em determinado período.

1.1 Objetivos

1.1.1 Objetivo Geral

- Realizar a predição da variação do valor das criptomoedas a partir dos dados históricos.

1.1.2 Objetivo Específico

- Estudar e analisar os dados históricos das criptomoedas.
- Estudar as técnicas de predição.
- Implementar as técnicas estudadas.

1.2 Organização do Texto

No Capítulo 2 estão descritas as técnicas estudadas mais relevantes ao projeto, enquanto no Capítulo 3 está descrita a metodologia a ser utilizada durante o desenvolvimento do trabalho. E por fim, no Capítulo 4 está descrita a proposta de trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos básicos, tipos e um breve histórico das criptomoedas, bem como as técnicas de aprendizado de máquina estudadas para realização da predição predição.

2.1 Criptomoedas

Criptomoedas são moedas virtuais, que usam a criptografia e a tecnologia chamada de blockchain para garantir o funcionamento descentralizado de negociações pela internet. São representadas por um código, protegido por criptografia e muito difícil de ser alterado. Por serem descentralizadas, elas podem ser transferidas de uma pessoa para outra sem a intermediação de um banco. Dessa forma, elas não possuem regulamentação do Banco Central. Seu único sistema de controle é a blockchain, um tipo de banco de dados, na qual cria-se um índice global para todas as transações dentro do mesmo mercado. É uma espécie de livro-razão, totalmente público e compartilhado. A ausência da mediação de terceiros cria o senso de confiança na comunicação direta entre as partes da transação.

A descrição do sistema eletrônico e anônimo de pagamento "B Money", foi publicado por Wei Dai (1998). Pouco tempo depois, NicK Szabo (1998) teorizou o "Bit Gold", exigindo uma função de prova de trabalho. A popularização das criptomoedas descentralizadas veio 11 anos depois, quando o pseudônimo Satoshi Nakamoto (2009) criou o Bitcoin utilizando-se de uma função criptográfica (SHA-256) como prova de trabalho. Desde então surgiram diversas moedas, a partir de 2014, surge a 2ª geração de criptomoedas, como Monero, Ethereum, entre outras. Essas criptomoedas possuem funcionalidades avançadas como endereços escondidos e contratos inteligentes. Em 2014, foi criada uma criptomoeda chamada RaiBlocks, resolvendo problemas contidos no Bitcoin, transações lentas e o alto consumo de energia. Ao contrário do Bitcoin, RaiBlocks é instantânea e não há taxas para efetuar transações. Em 2018, a moeda foi renomeada para Nano.

2.1.1 Bitcoin

A primeira criptomoeda que surgiu foi o Bitcoin, criada em 2008 pelo pseudônimo de Satoshi Nakamoto (2009). Ele publicou um artigo chamado “Bitcoin: a peer-to-peer electronic cash system”, detalhando o funcionamento da criptomoeda e em janeiro de 2009 o sistema foi colocado no ar pela primeira vez.

2.1.2 Ethereum

A Ethereum é uma plataforma de código aberto para transações. Utiliza o Ether, sua moeda virtual. Ela é considerada a segunda moeda de mais valor entre as moedas virtuais comercializadas, perdendo apenas para o BitCoin. É também a mais conhecida além do mais famosa moeda virtual. Em 2016, problemas internos fizeram com que ela se quebrasse em duas, formando além dela mesma, mais uma versão clássica.

2.1.3 Ethereum Classic

A Ethereum Classic surgiu recentemente, ainda em 2016, quando um ataque de hackers ao sistema utilizado pela Ethereum fez com que desenvolvedores e mineiros da moeda escolhessem caminhos diferentes.

Apesar de parecida com a Ethereum, a Ethereum Classic é uma versão diferente, e portanto, uma moeda virtual diferente. É uma continuação da plataforma original da Ethereum, por isso "Classic".

2.1.4 Monero

Criada em 2014, a moeda virtual pretende através tornar as transações o mais anônimas possível, além tornar a mineração mais igualitária, para que mais pessoas possam ter acesso a esse tipo de operação. Seu formato fez com que fosse adotada por sites como o darknet Alphabay, que chegou a ter mais de 200 mil usuários. O site foi fechado pela justiça em 2017 devido à ilegalidade de seu funcionamento, vendendo coisas como contas roubadas da Uber.

2.1.5 Litecoin

A moeda virtual surgiu em outubro de 2011 e é considerada mais leve para processamento do que o Bitcoin. A grande da atração dessa moeda é a possibilidade de mineração utilizando hardwares mais modestos, sem máquinas maiores, por exemplo. Com isso, ela se propõe uma moeda mais democrática e fácil de ser utilizada. Apesar desse aspecto, é uma das 10 mais valiosas moedas em negociações no mundo financeiro virtual.

2.2 Técnicas Aprendizado de Máquina

Um sub-campo da inteligência artificial, considerado um campo da ciência computacional, o aprendizado de máquina possui, tem como objetivo principal desenvolver métodos eficientes de reconhecimento de padrões que serão capazes de generalizar além dos exemplos do conjunto de treinamento, para que possam de acordo com problema em questão e os dados disponíveis para análise obterem bons resultados (ROZA, 2016). O aprendizado de máquina pode ser dividido em dois grupos de algoritmos: *aprendizagem não-supervisionada* e *aprendizagem supervisionada*. De modo geral, a aprendizagem não-supervisionada, não utiliza informações das variáveis de saída. Os dados de entradas são analisados e agrupados conforme a proximidade dos seus valores. Para cada grupo de registros é utilizado um rótulo de identificação. Enquanto na aprendizagem supervisionada, deve existir um supervisor, que é dado pelo registro dos valores das variáveis de saída, que são as variáveis que se deseja predizer a partir dos dados existentes. Como resultado, obtém-se um modelo que descreva o conjunto de dados utilizados e espera-se que ele permita prever o comportamento da saída para novas entradas (ROZA, 2016).

Regressão é uma dessas técnicas, que permite investigar e compreender a relação entre uma variável de resposta e variáveis específicas, através da construção de um modelo. A análise da regressão pode ser usada como um método descritivo da análise de dados com vários objetivos, como descrever a relação entre variáveis para entender um processo, prever o valor de uma variável a partir do valor das outras variáveis, substituir a medição de uma variável pela observação dos valores de outras variáveis, e ainda controlar os valores de uma variável em uma faixa de interesse.

2.2.1 Regressão Linear Simples

A regressão linear simples é a estimativa do mínimo quadrado de um modelo de regressão linear com uma única variável de resposta. Em outras palavras, regressão linear simples é uma linha reta através do conjunto de pontos n de tal forma que faça a soma dos quadrados residuais do modelo tão pequena quanto (AMARAL, 2016). A inclinação da reta é igual à correlação entre a variável de resposta Y_i e a variável específica X_i, seja corrigida pela relação de desvios padrão destas variáveis. A interseção da linha

reta é tal que passa pelo centro de massa (x,y) dos pontos de dados. Sendo assim, temos que a Equação 2.1 é a aproximação por regressão simples.

$$\mathrm{Q}(\alpha, \beta) = \sum_{i=1}^{\infty}(y_i - \alpha x_i - \beta)^2 \quad (2.1)$$

Quando se tem mais de uma variável específica, chamamos de regressão linear múltipla.

2.2.2 Regressão Logística

O modelo de regressão logística é semelhante ao modelo de regressão linear. No entanto, no modelo logístico a variável resposta Y_i é binária. Uma variável binária assume dois valores, como por exemplo, $Y_i = 0$ e $Y_i = 1$, onde $Y_i = 1$ é o evento de interesse, denominado "sucesso". A variável resposta Y tem distribuição Bernoulli $(1, \pi)$, com probabilidade de sucesso $P(Y_i = 1) = \pi_i$ e de fracasso $P(Y_i = 0) = 1 - \pi_i$. Sendo assim, temos na Equação 2.2 a aproximação por regressão logística.

$$\mathrm{E}(\mathrm{Y}_i) = \pi_i = \alpha x_i + \beta \quad (2.2)$$

Assim como no modelo de regressão linear simples, um modelo de regressão logística múltipla é usado para o caso de regressão com mais de uma variável de específica.

2.2.3 Regressão Bayesiana

Os métodos bayesianos são técnicas alternativas aos métodos clássicos de inferência. Enquanto nos modelos clássicos de inferência, tem-se o parâmetro desconhecido considerado uma constante fixa, a informação amostral é a única considerada (através da equação de verossimilhança) e não se estima probabilidade para intervalos de confiança. No modelo de inferência bayesiano o parâmetro desconhecido é considerado uma variável aleatória que segue uma distribuição a priori, e assim sendo, podem-se considerar informações de estudos anteriores, conhecimento pessoal, etc. E pode-se também estimar a probabilidade para intervalos de confiança. Sendo assim, podemos assumir a estatística bayesiana como um processo de diminuição de incerteza sobre o desconhecido que se baseia em dados estatísticos e em evidências prévias (MANCUSO, 2010).

A ideia da inferência bayesiana é combinar a informação a priori com a informação proveniente dos dados amostrais, ou seja, combinar a distribuição a priori e a função de verossimilhança. Esta combinação é feita através do “Teorema de Bayes” , originando a distribuição “a posteriori”. Para o

cálculo da probabilidade de um evento A dado que um evento B ocorreu, pelo Teorema de Bayes temos a Equação 2.3.

$$P(A|B) = P(B|A).P(A)\frac{}{P(B)} \quad (2.3)$$

2.2.4 Random Forest

Floresta aleatória (*Random Forest*), de modo simples, é um método que cria várias árvores de decisão e as combina para obter uma predição com maior precisão e estabilidade, podendo ser usado tanto para classificação como para regressão. Seu funcionamento ocorre através da definição de um número de árvores de decisão, sendo cada uma delas diferente uma das outras e responsável por um conjunto de regras distinto (DONGES, 2018). Quando houver um dado de entrada, todas as árvores, através de uma seleção randômica de nós, classificam o dado e disponibilizam seus resultados para que a classificação final seja definida. Sendo assim, o resultado se dá pela moda estatística das categorias de todos os resultados ou a média numérica dos mesmos (SOUZA, 2017). Na Figura 1, pode-se ver uma floresta aleatória genérica com n árvores.

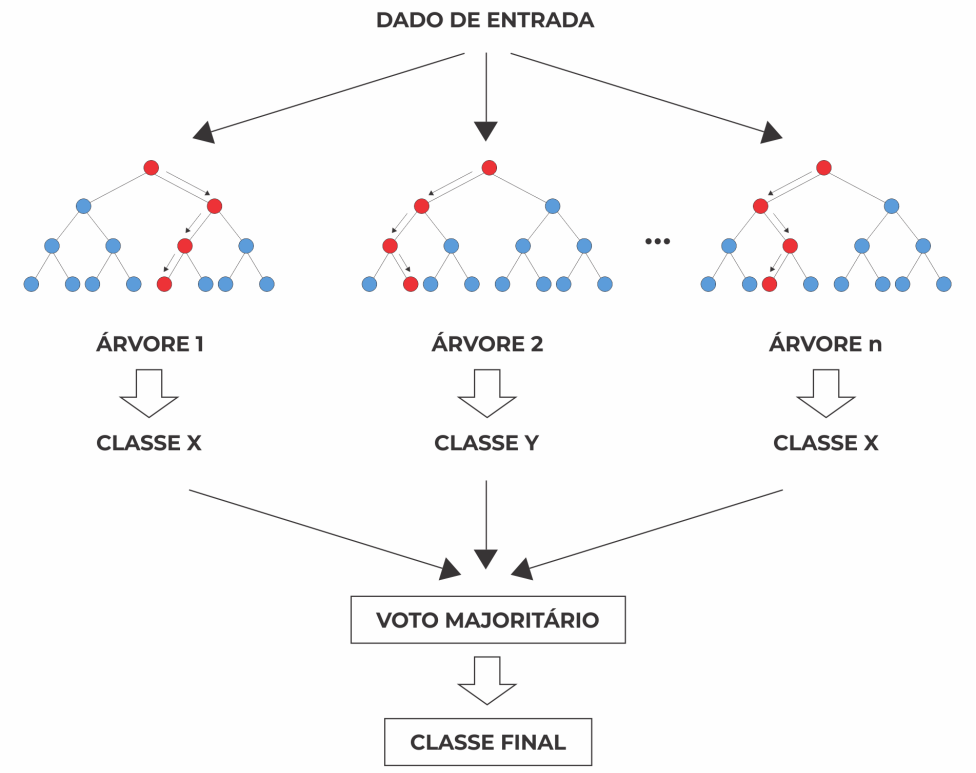

Figura 1 – Funcionamento do algoritmo Random Forest

Fonte: Adaptação de (DIMITRIADIS; LIPARAS, 2018)

3 METODOLOGIA

3.1 Base de Dados

Para nossos estudos utilizaremos uma base de dados publicada na comunidade Kaggle por Rajkumar (2018). O *dataset* escolhido tem um arquivo (.csv) para cada moeda. O histórico de preços está disponível diariamente a partir de 28 de abril de 2013. Esse conjunto de dados tem as informações históricas sobre preços de algumas das principais moedas de criptografia, são elas:

- Bitcoin
- Ethereum
- Ripple
- Monero
- Litecoin

Este conjunto de dados oferece 7 (sete) informações muito importantes para serem analisadas:

- Data de observação.
- Preço de abertura no dia indicado.
- Maior preço no dia indicado.
- Menor preço no dia indicado.
- Preço de fechamento no dia indicado.
- Volume de transações no dia indicado.
- Valor de mercado em Dolar (USD).

Sobre a comunidade Kaggle, é uma comunidade online, comprada em 2017 pela Google Inc, que permite aos usuários encontrar e publicar conjuntos de dados (datasets), explorar e construir modelos em um ambiente de ciência de dados baseado na Web, trabalhar com outros cientistas de dados, aprender e participar de competições para resolver os desafios da ciência de dados. A Kaggle começou oferecendo competições de aprendizado de máquina e agora, além de outros serviços, oferece também uma plataforma de *datasets* públicos.

3.2 Aparato Técnico

Para este estudo, precisa-se de um aparato técnico mínimo que incluem os conjuntos de dados já mencionados na seção anterior e um ambiente de programação onde possamos importar este conjunto de dados e tratá-los da maneira correta para obtermos resultados satisfatórios. Escolheu-se o ambiente R e a linguagem R para tal tarefa, é muito útil e serão utilizados alguns pacotes do repositório do R para facilitar a manipulação dos dados.

3.2.1 Linguagem R

R é uma linguagem e ambiente para computação estatística e gráficos. É um projeto GNU desenvolvido por Ross Ihaka e Robert Gentleman (1993), que fornece uma ampla variedade de estatística (modelagem linear e não linear, testes estatísticos, análise de séries temporais, classificação, agrupamento, entre outras) e técnicas gráficas, além de ser altamente extensível com o uso dos pacotes, que são bibliotecas para sub-rotinas específicas ou áreas de estudo específicas. Um conjunto de pacotes é incluído com a instalação de R, mas existem muitos outros em um repositório chamado de CRAN (do inglês, Comprehensive R Archive Network). Um dos pontos fortes de R é a facilidade com que plotagens bem projetadas podem ser produzidas, incluindo símbolos matemáticos e fórmulas. Além da linguagem R, utiliza-se opcionalmente para maior produtividade uma interface gráfica chamada RStudio, que facilita a programação e utilização das ferramentas (R Project, 1993).

4 PROPOSTA

Com base na proposta de Rameshbabu (2017), de estudar técnicas de predição de preços de bitcoins, decidiu-se realizar a predição de diferentes tipos de criptomoedas utilizando-se dos algoritmos e técnicas de regressão, análise e predição dos dados apresentadas no Capítulo 2 e Capítulo 3.

Antes de mais nada, é preciso estudar as tendências e periodicidades envolvidas na variação do preço das criptomoedas e com base nessas informações decidir qual modelo de predição deve ser utilizado. Sendo assim, primeiro deve-se realizar um estudo das diversas técnicas de predição com a finalidade de compreender quais são melhores para cada conjunto de dados.

Dessa forma, por meio da linguagem R, deve-se utilizar o conjunto de dados referente aos valores históricos das criptomoedas que desejamos, para conhecer a estrutura dos dados nele contidos. Uma vez estudada a estrutura dos dados, separamos os dados em dois conjuntos. Um para treinamento, no qual temos as informações de valor de abertura, menor e maior valor, valor de fechamento e ainda, volume de transações, volume de moedas circulantes e data, que está disponível no repositório Cryptocurrency Historical Prices (2018). E um segundo conjunto de dados de teste, que servem para comparação, sendo utilizados no modelo proposto para verificar a tendência estimada pela predição com o que de fato ocorreu. O conjunto de treinamento tem data anterior ao conjunto de teste.

Com base nesse conjunto adquirido, pode-se então selecionar os modelos de predição que melhor se adaptam aos dados que desejamos tratar, realizando assim os testes com o conjunto de treinamento. Desta forma, será apresentada uma comparação de técnicas de aprendizagem de máquina para análise de dados, que sejam eficientes na predição da variação de preço de criptomoedas. Bem como um breve estudo sobre a utilização de agentes inteligentes que sejam capazes de avaliar de forma independente a predição variação de preços de diferentes criptomoedas, elegendo as mais rentáveis segundo as tendências.

4.1 Cronograma

Tabela 1 – Cronograma das atividades previstas

Etapa	Mês												
	07	08	09	10	11	12	01	02	03	04	05	06	07
E1	*X*	*X*	*X*										
E2				*X*	*X*	*X*							
E3						*X*	*X*	*X*	*X*	*X*			
E4										*X*	*X*	*X*	
E5													*X*

- **E1 -** Estudo das técnicas de análise de dados e aprendizado de máquina.
- **E2 -** Estudos sobre a estrutura do conjunto de dados.
- **E3 -** Definição das técnicas escolhidas e implementação dos modelos.
- **E4 -** Análise das tendências preditas, conclusões e elaboração do documento final.
- **E5 -** Apresentação.

REFERÊNCIAS

AMARAL, F. *Introdução à Ciência de Dados. Mineração de Dados e Big Data (Em Portuguese do Brasil)*. [S.l.]: Alta Books, 2016. ISBN 8576089343. Citado na página 10.

BARBOSA, J. P. *Os três pilares da inovação – Machine Learning, Big Data e IoT*. 2017. Acessado em: 27 out. 2018. Disponível em: <http://igti.com.br/blog/os-tres-pilares-da-inovacao-machine-learning-big-data-e-iot/>. Citado na página 7.

CARLOS, E. *Afinal, o que é mineração?* 2018. Acessado em: 27 Out. 2018. Disponível em: <https://www.criptomoedasfacil.com/afinal-o-que-e-mineracao/>. Citado na página 7.

Cryptocurrency Historical Prices. 2018. Acessado em: 08 Dez. 2018. Disponível em: <https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory>. Citado na página 15.

DIMITRIADIS, S. I.; LIPARAS, D. *How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database*. 2018. Acessado em: 06 Dez. 2018. Disponível em: <http://www.nrronline.org/temp/NeuralRegenRes136962-43545_120544.pdf>. Citado na página 12.

DONGES, N. *The Random Forest Algorithm*. 2018. Acessado em: 09 Dez. 2018. Disponível em: <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>. Citado na página 12.

MANCUSO, A. C. B. *Métodos Bayesianos em Metanálise*. 2010. Acessado em: 04 Dez. 2018. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/29108/000775678.pdf?sequence=1>. Citado na página 11.

NicK Szabo. 1998. Acessado em: 27 Out. 2018. Disponível em: <https://coincentral.com/who-is-nick-szabo/>. Citado na página 9.

R Project. *What is R?* 1993. Acessado em: 4 Dez. 2018. Disponível em: <https://www.r-project.org/about.html>. Citado na página 14.

RAJKUMAR, S. *Dataset: Cryptocurrency Historical Prices*. 2018. Acessado em: 08 Dez. 2018. Disponível em: <https://www.kaggle.com/sudalairajkumar>. Citado na página 13.

RAMESHBABU, A. *Forecasting of Bitcoin Prices*. 2017. Acessado em: 18 Nov. 2018. Disponível em: <https://www.kaggle.com/ara0303/forecasting-of-bitcoin-prices>. Citado na página 15.

Ross Ihaka; Robert Gentleman. 1993. Acessado em: 4 Dez. 2018. Citado na página 14.

ROZA, F. S. da. *Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas*. 2016. Acessado em: 09 Dez. 2018. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/171569/PFC_2016-1%20Felippe_Roza.pdf?sequence=1>. Citado na página 10.

RUSSEL, S.; NORVIG, P. *Inteligência Artificial*. [S.l.]: Elsevier, 2013. ISBN 8535237011. Citado na página 7.

Satoshi Nakamoto. 2009. Acessado em: 27 Out. 2018. Disponível em: <https://en.bitcoin.it/wiki/Satoshi_Nakamoto>. Citado na página 9.

SCHIAVON, G. *Por que o valor do bitcoin varia tanto?* 2017. Acessado em: 27 Out. 2018. Disponível em: <https://foxbit.com.br/blog/por-que-o-valor-do-bitcoin-varia-tanto-descubra/>. Citado na página 7.

SOUZA, L. A. M. de. *Aplicação de Aprendizado de Máquina para Predição de Prioridade em Gestão de Incidentes*. 2017. Acessado em: 09 Dez. 2018. Disponível em: <http://bsi.uniriotec.br/tcc/textos/201712LucasSouza.pdf>. Citado na página 12.

Wei Dai. 1998. Acessado em: 27 Out. 2018. Disponível em: <https://en.bitcoin.it/wiki/Wei_Dai>. Citado na página 9.