ForeSeries - Uma abordagem para predição de valores em séries temporais

Faculdade de Tecnologia de São José do Rio Preto

Johny William de Oliveira Alves, Josiane Rodrigues de Souza Nakagawa, Sergio Ricardo Borges Junior

e-mail: contato@johnywalves.com.br, josinakagawa@gmail.com, sergio@fatecriopreto.edu.br

Resumo: Neste estudo, foram apresentados conceitos sobre a utilização de ciência de dados, ferramentas de estatísticas, tecnologias, bibliotecas e funções na linguagem R, explorando a aplicação prática destes conceitos em fontes de dados de acesso público, como as cotações da Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&FBOVESPA S.A.), a fim de realizar predição de valores em séries temporais. Através uso de regressão linear múltipla, árvore de decisão, floresta aleatória e redes neurais com funções disponíveis no ambiente da linguagem R, foi demonstrada uma metodologia para a predição de valores em séries temporais pela comparação de resultados obtidos apontando o melhor algoritmo para cada análise. O método proposto pode ser aplicado para estimativa de demandas, valores de mercadorias, necessidades sociais entre outras informações futuras de movimentações de pessoas e valores.
Palavras-chave: Análise de Dados, Predição, Séries Temporais

Abstract: *In this study we present concepts about the use of data science, statistical tools, technologies, libraries and functions in the R language, exploring the practical application of these concepts in data sources of public access, such as the Stock Exchange, Commodities and Futures of São Paulo (BM&F BOVESPA SA), in order to predict values in time series. Through the use of linear regression, decision tree, random forest and neural networks with functions available in the environment of the R language, showing the methodology for the prediction of values in time series by comparing obtained results indicating the best algorithm for each analysis. The presented method can be applied to estimate demands, values of goods, social needs among other future information of movements of people and values.*
Keywords: *Data analysis, Prediction, Time Series*

1. Introdução

Atualmente, nossa sociedade gera um grande volume de dados por meio de planilhas, mídias sociais, entre outras fontes de dados, assim, um dos grandes desafios das empresas é filtrar esses dados com intuito de compreende-los e gerar uma informação útil. Entende-se que profissionais com visão lógica e analítica são capazes de aplicar conceitos teóricos e metodologias inovadoras de análise de dados para gerar ganho de valor.

No entanto é necessário analisar os índices e resultados em séries temporais para que a empresa possa ter vantagens competitivas, podemos citar exemplos tais como: o Wal-Mart descobrir que deve reforçar seus estoques de Pop-Tarts de morango antes de um furacão (MAYER-SCHONBERGER e CUKIER, 2013) e (PROVOST e FAWCETT, 2016), ou do Target

que foi capaz de descobrir que uma jovem estava grávida antes de toda sua família (DUHIGG, 2012).

Com base neste contexto, o presente trabalho tem por objetivo propor uma abordagem para predição de valores em séries temporais por meio de funções estatísticas e ferramentas computacionais, com intuito de compreender valores futuros como demanda de mercadorias, necessidades sociais e valores de bolsa de valores.

O presente trabalho está organizado da seguinte maneira. A Seção 2 descreve a metodologia utilizado no desenvolvimento do trabalho. A seção 3 descreve fundamentação teórica base para a aplicação das teóricas. Na seção 4 apresenta a exemplos de aplicação da análise de dados em séries temporais. A seção 5 descreve o desenvolvimento da abordagem ForeSeries com o estudo de caso das ações negociadas na Bovespa. A conclusão é apresentada na seção 6.

2. Metodologia

O presente trabalho possui caráter exploratório, pois busca identificar os principais critérios de análise de um conjunto de dados, utilizando conceitos estatísticos aplicados em algoritmos de predição combinados com linguagem R. Na fundamentação teórica serão utilizados trabalhos de conclusão de curso, dissertações, teses, livros e artigos científicos. O plano de trabalho foi dividido em 4 etapas, conforme a seguir:

Etapa 1: Levantamento bibliográfico

O levantamento bibliográfico será realizado a partir da leitura de trabalhos relacionados ao tema de predição de valores, séries temporais e o processo de descoberta de conhecimento. Com isso, a presente etapa permitiu a escolha das tarefas e algoritmos de árvores de regressão a serem utilizados.

Etapa 2: Aplicações do Mundo Real

Essa etapa permitiu investigar alguns estudos de caso e entender a relevância do conhecimento em análise de dados, bem como as abordagens utilizadas e os respectivos algoritmos empregados.

Etapa 3: Investigação de algoritmos de predição de valores

Essa etapa permitiu investigar como funcionam os alguns algoritmos de predição de valores disponíveis na ferramenta R, destacando-se os algoritmos de regressão linear múltipla, árvore de decisão, floresta aleatória e redes neurais. Entretanto, apesar de ser mais flexível, foram obtidos melhores resultados para predição de valores em séries temporais, com base no conjunto de dados quando aplicado floresta aleatória.

Etapa 4: Estudo de Casos

Com os dados dos valores de fechamento das ações de participação negociadas pelo Bovespa, a partir da limpeza e organização dos dados foram aplicados treino, validação e testes para gerar dados de verificação de comportamento e predição de valores em conjunto a plotagem mostrando de forma gráfica os resultados das análises.

A seguir, será apresentada a fundamentação teórica necessária para o desenvolvimento do presente projeto.

Por fim, ressalta-se a apuração dos resultados.

3. Fundamentação Teórica

A fundamentação teórica deste trabalho aborda os principais conceitos utilizados no desenvolvimento, com destaque para: correlação, regressão linear múltipla, aprendizado de máquina, árvore de decisão, floresta aleatória e rede neural o quais serão descritos a seguir.

3.1 Correlação

A principal característica de correlação é quando duas variáveis estão relacionadas e representam dados quantitativos. Segundo Triola (2013) “O coeficiente de correlação linear r mede a força da correlação entre os valores quantitativos emparelhados de x e y em uma amostra.”, foi utilizado o coeficiente de correlação de Pearson.

Para estudos estatísticos a amostra de uma população é analisada para entender a relação entre duas variáveis através da covariância e coeficiente de correlação, “A correlação indica se, e com que intensidade, os valores de uma variável aumentam (ou diminuem)” (SILVA, PERES E BOSCARIOLI, 2016).

Pautado nesses conceitos foi utilizado o método de Pearson pela bibliografia vasta e de fácil acesso. No método de Person é aplicada a fórmula em que a (covariância é dividida pelo produto do desvio padrão de X e Y) conforme Equação (1), que foi adaptada de Guimarães (2017).

$$r = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

O coeficiente varia de -1 a +1, quanto mais próximo do extremo demonstra a relação de significância, mas sempre se deve estar atento para não confundir correlação com causalidade.

No presente trabalho a correlação foi aplicada para analisar quais atributos possuem ligação entre si, para facilitar a criação das fórmulas para geração dos modelos de predição.

3.2 Regressão Linear Múltipla

A regressão linear múltipla representa a proporção da variação em Y que é explicada pelo conjunto de variáveis independentes" (LEVINE et al., 2012) e, ainda, segundo Provost e Fawcett (2016), para regressão ou estimativa de valor tenta-se predizer o valor numérico de uma variável, prevendo na classificação se alguma coisa vai acontecer e na regressão o quanto essa coisa vai acontecer.

Ainda, Silva et al. (2016) demonstra que a análise preditiva busca descobrir o relacionamento de um determinado conjunto de dados característicos (atributos descritivos) e rótulos associados (atributos de classe), por exemplo: rótulos de identificação (nome de vinhos), o método de classificação faz uma apresentação de rótulos ou predição categórica, enquanto a regressão trata-se de uma predição numérica ou seja, dados numéricos.

No presente trabalho, a regressão linear múltipla foi utilizada para determinar o quanto os resultados foram preditos de maneira correta. Em outras palavras, buscou-se determinar a acurácia dos resultados obtidos, ou seja, o quanto eles foram corretos.

3.3 Aprendizado de máquina (*Machine Learning*)

O aprendizado de máquina ou *machine learning* pode ser explicada como "A coleta de métodos para a extração (previsão) de modelos a partir de dados, agora conhecida como métodos de aprendizado de máquina" (PROVOST e FAWCETT, 2016).

Conforme as aplicabilidades de aprendizado de máquina são subdivididas em três vertentes com vários algoritmos já desenvolvidos para a implementação destes conceitos, alguns exemplos apresentados por Alcântara (2017).

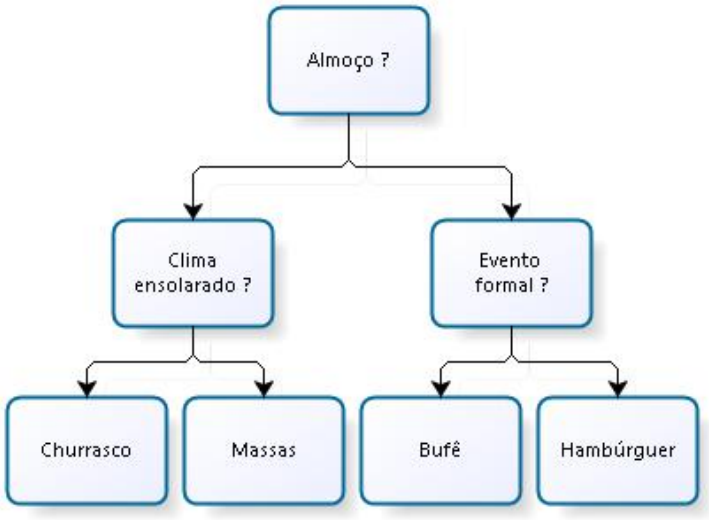
- **Classificação**: *Bayes, Decision Tree, Random Forest, Neural Network, K-Neast Neighbors*…;
- **Regressão**: *Polinomial*, *Linear*, *Logistic*…;
- **Análise de *Cluster***: *Hierarchical*, *Density Based, Expectation Maximization, K-Mean, K-Median, Density Based*...

Nesta abordagem foi aplicado para predição de valores os modelos de árvore de decisão, floresta aleatória e redes neurais.

3.3.1 Árvore de Decisão (*Decision Tree*)

Árvore de decisão são estruturas de escolha que organizam caminhos de gerado pelas respostas afirmavas e negativas para determinadas perguntas, normalmente são modelos supervisionados de valores discretos utilizados em um conjunto de dados, conforme se observa na Figura 1.1.

Figura 1.1 - Exemplo de uma árvore de decisão apresentando os possíveis caminhos para encontrar o resultado desejado, com o caminho da esquerda apontando a resposta afirmativa para as questões.



Fonte: elaborado pelos autores

No presente trabalho utilizou-se a biblioteca *RPart* (*Recursive Partition and Regression Trees*), Therneau et al. (2017), que faz uso do modelo *Classification and Regression Trees (CART)* que classifica os valores para busca dos melhores índices de acerto.

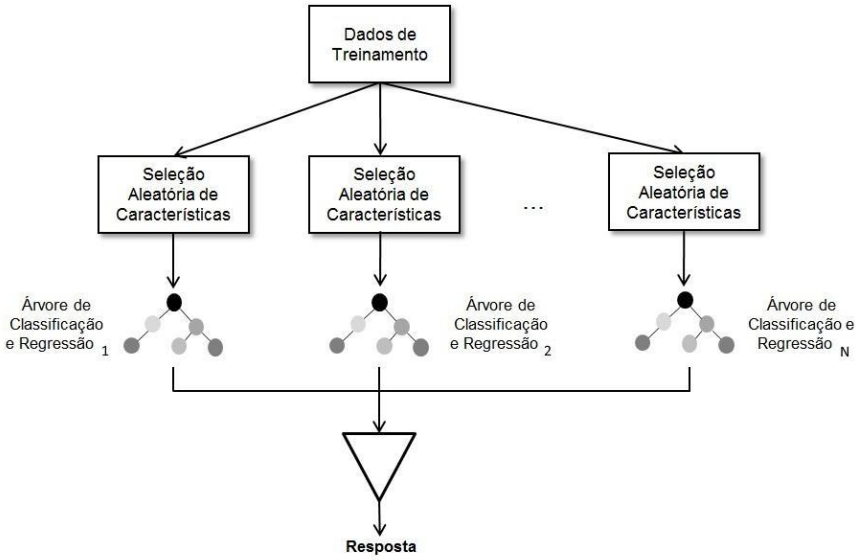
3.3.2 Floresta Aleatória (*Random Forest*)

O algoritmo de uma floresta aleatória é construído com várias árvores com entrada de valores aleatórios, onde o resultado é definido pelo maior número de ocorrências.

> O método random forest (floresta aleatória) foi desenvolvido por Breiman (2001) para combinar vários modelos, especificamente árvore de classificação e/ou regressão, com base em vetores de características (covariáveis), os quais são gerados de maneira aleatória e independente a partir do conjunto de dados. (BORGES, 2016)

Entende-se que quando aplicado em regressão cada árvore produz uma resposta numérica com valores preditos aleatórios, com uma construção aleatória o aprendizado de máquina seleciona as melhores e refinando seu modelo como apresentado na Figura 2.1.

Figura 2.1 - Exemplo de desenvolvimento de uma *Random Forest*.



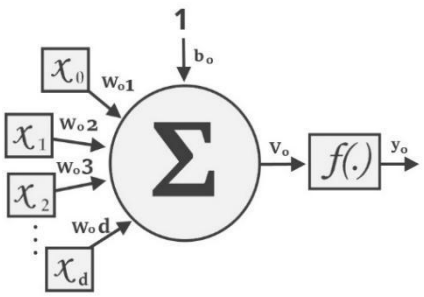
Fonte: Borges Jr. et al. (2016).

Nesse estudo foi utilizado a biblioteca *randomForest*, Liaw e Wiener (2015), para a linguagem R.

3.3.3 Rede Neural (*Neural Network*)

Rede Neural Artificial, ou somente Rede Neural, é um campo da ciência que estuda tipos de classificação por meio de sinapses de um neurônio artificial. Segundo Silva et al. (2016) “o neurônio artificial é, na realidade, uma função que mapeia entradas e saídas.”. A Figura 3.1 ilustra uma simulação de funcionamento de um neurônio no modelo *perceptron* com valores de entrada $\{x_0, x_1, x_2 \dots x_d\}$ sendo multiplicados por pesos sinápticos $\{w_{01}, w_{02}, w_{03} \dots w_{0d}\}$. Esses valores são somados (somatório) e acrescidos por um valor externo (b_0), o qual funciona como sendo um peso. O valor resultante, ou melhor, valor de saída (y_0), é obtido por meio da função f(.), com base no valor resultado do somatório (v_0). A função f(.) um classificador binário definindo se o resultado é retorno 0 ou 1.

Figura 3.1 – A estrutura de um neurônio artificial do tipo *Perceptron*.



Fonte: adaptado de Silva et al. (2016).

No processo de aprendizado de máquina o sistema estipula e testa os valores dos pesos para encontrar o resultado mais próximo do resultado esperado.

O modelo *perception* quando encadeado e organizado em várias camadas é chamado de *perception* multicamadas, com a função classificadora com um espectro entre 0 e 1.

No presente trabalho utilizou-se a função *neuralnet*, com o modelo *perception* de multicamadas utilizando o algoritmo do *backpropagation*, que possui um padrão de descoberta inversa, com os valores sendo testados a partir da saída.

Entre as principais vantagens de redes neurais pode-se citar detecção de fraudes, reconhecimento de digitais, predição de mercado financeiro entre outros. A escolha de redes neurais para ser aplicado no presente trabalho foi relevante devido a flexibilidade de trabalhar com valores flutuantes, comuns em algumas séries temporais.

4. Aplicação do mundo real

O cenário atual mostra que grandes empresas estão aplicando tecnologias para melhoramento genético, descoberta de novas doenças, percepção de mercado financeiro, desenvolvimento de novos produtos, criar aplicativos moveis, visando entender o comportamento do cliente para satisfazer as demandas do mercado consumidor.

Entretanto, a análise de dados é um dos fatores mais relevantes para tomada de decisões avaliando os riscos e visando maximizar os resultados. Assim, a seguir são descritos 3 estudos de casos, os quais foram descritos por apresentarem aplicações diferenciadas entre si.

4.1 Entendendo o cliente

Charles Duhigg, repórter e escritor não ficcional que trabalha para o The New York Times, descreve no seu livro O poder do hábito (2012) como entender o comportamento de um consumidor “Você coleta dados. Quantidades enormes, quase inconcebíveis de dados” descrevendo o sucesso da Target ao analisar os dados dos clientes podendo identificá-los pelo cartão de crédito emitido pela Target, uso de um cartão de fidelidade, um cupom de desconto entregue pelo correio, entre outras maneiras dando mais valor para os dados:

> Para um leigo, dois consumidores que compram suco de laranja parecem iguais. É preciso um tipo especial de matemático para se dar conta de que um deles é uma mulher de 34 anos comprando suco para os filhos (e por isso talvez goste de receber um cupom para um DVD infantil) e outro é um homem solteiro de 28 anos que bebe suco depois de sair para correr (e assim talvez interessado em descontos de tênis). (DUHIGG, 2012)

A Target entendeu que poderia oferecer cupons de desconto para consumidores, folhetos diferentes pelos correios, anúncios por e-mail podendo chegar a entregar um livro de descontos e ter certeza que o cliente compraria, outras empresas que também fizeram isso são a Amazon.com, Best Buy, Hewlett-Packard, Capital One entre outros.

Em alguns estados brasileiros o governo estadual adotou uma política de retornar parte do valor arrecadado pelo Imposto sobre Circulação de Mercadorias e Serviços (ICMS) para incentivar uma fiscalização cidadã, exemplo a lei 12.685 de 28 de agosto de 2007 da Secretaria da Fazenda do Estado de São Paulo e o decreto 50.199 de 4 de abril de 2013 da Secretaria da Fazenda do Rio Grande do Sul, para poder usufruir deste benefício o cliente deve se identificar fornecendo o documento de Cadastro de Pessoa Física (CPF) ou o Cadastro Nacional de Pessoa Jurídica (CNPJ), uma ótima maneira de identificar o comprador para otimizar a relação, possibilitando a criação de um histórico de compras, gerando dados para compreender o comportamento de consumo e a predição de itens que podem fazer parte de próximas compras, que pode ser usado para oferecer novos produtos ou organizar a disposição dos produtos para maximizar o resultado da operação.

4.2 Prevendo o futuro

Mercado de ações e valores e mercados futuros é um mar de incertezas e especulações, pessoas abrem mão de valores que dispõe atualmente na crença poderem dispor de mais em um determinado futuro, as ações estão atreladas a capacidade da empresa de prover as modalidades de rendimento e seu valor é determinado por essa percepção baseado em preços históricos, assim como a análise disponibilizada pelas corretoras destes valores, reações a medidas do governo como fixação de determinadas taxas, possibilidade ganhos pelo lançamento de novos produtos entre outros.

Por exemplo do banco norte-americano J.P. Morgan que faz uso de uma complexa estrutura de tecnologia para efetuar a compra e venda de ações no mercado financeiro. Os sistemas de *Big Data* interpretam milhares de informações para prever tendências e tomar a melhor decisão na hora de comprar ou vender ações, relacionando acontecimentos políticos e econômicos com as transações da Bolsa de Valores.

4.3 Plantando melhor

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) pioneira em pesquisa na área, destaca a importância de aplicações de análise de dados nos agronegócios para aumentar a produtividade, de acordo com a Embrapa neste ano o XI Congresso Brasileiro de Agroinformática (SBIAgro, 2017), com o objetivo de compartilhar resultados de pesquisa traz o tema “Ciência de Dados na Era da Agricultura Digital”.

A demanda crescente de desenvolvimento sustentável e maximização de resultados são fatores que impulsionam as empresas a buscar tecnologias eficientes para gestão de recursos da cadeia de suprimentos, segundo Maia Junior (2014) ”De 2012 para cá, a Monsanto desembolsou 1,2 bilhão de dólares para comprar duas empresas que desenvolvem tecnologias na área, a Precision Planting, que produz equipamentos e software, e a The Climate, cujo negócio é fazer previsões meteorológicas.”.

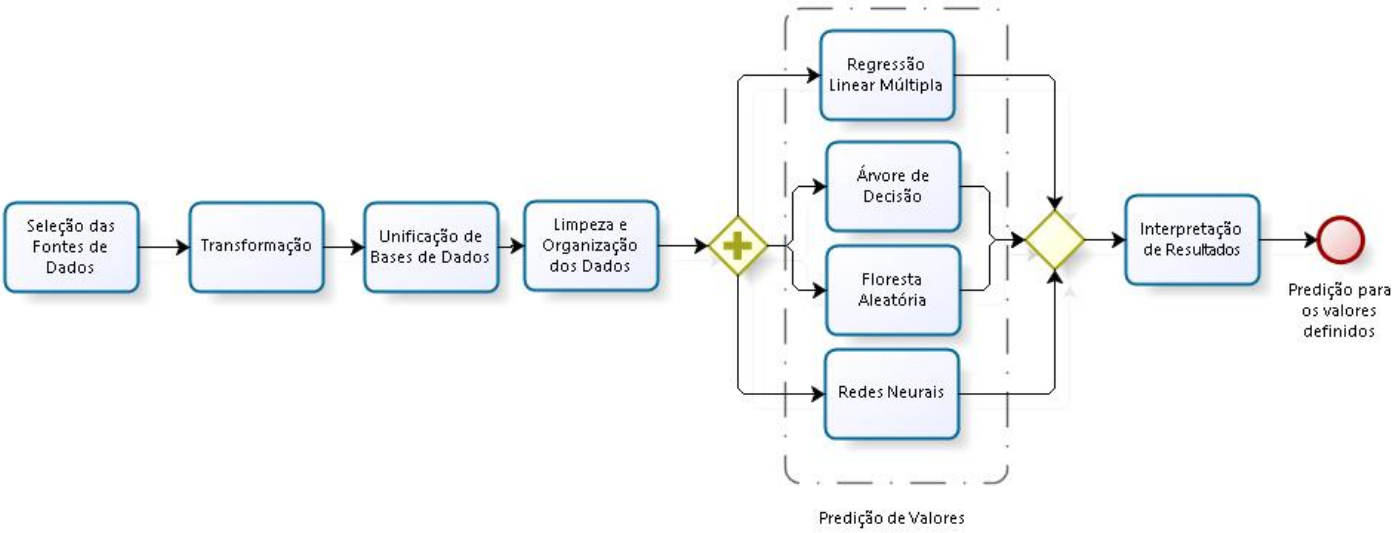
Um dos maiores gargalos é a falta de profissionais com visão lógica e analítica para compreender esse grande volume de dados e trazer informações importantes para tomada de decisões automatizadas, neste sentido entende-se a necessidade de desenvolver metodologias de processamento de dados para minimização de custos trazendo as melhores soluções para as empresas competir no cenário multinacional.

5. Desenvolvimento

O presente trabalho descreve uma abordagem para predição de valores em séries temporais que permite comparar as estimativas resultantes de quatro métodos distintos: regressão linear múltipla, árvore de decisão (*RPART*), florestas aleatórias e redes neurais como exemplo de modelos de predição.

A Figura 4.1 descreve o fluxo de trabalho da abordagem proposta, denominada “ForeSeries”, a qual permite efetuar predição em séries temporais. Observa-se que abordagem ForeSeries possui seis etapas: “Seleção de Fontes de Dados”, “Transformação”, “Unificação de Bases de Dados”, “Limpeza e Organização dos Dados”, “Predição de Valores” e “Interpretação de Resultados”.

Figura 4.1 - Fluxo de trabalho da abordagem ForeSeries.



Fonte: elaborado pelos autores pelo Bizagi Modeler.

A etapa de “Seleção de Fontes de Dados” permite a seleção de uma série temporal a qual será manipulada para a predição de valores. Ressalta-se que podem ser selecionadas mais do que uma fonte de dados. Por exemplo, duas planilhas com dados temporais obtidas de fontes diferentes. A etapa “Transformação” provê a conversão das bases de dados selecionadas em um formato adequado para ser utilizado na Linguagem R, que neste projeto, é o formato CSV. A escolha da linguagem R foi devida a sua flexibilidade, facilidade e gratuidade.

A partir das bases já transformadas, ou seja, já em formato CSV, a etapa de “Unificação de Bases de Dados” permite reuni-las em um único conjunto de dados por meio da ocorrência da data das instâncias (registros). Em seguida, a etapa de “Limpeza e Organização dos Dados” remove informações eventualmente irrelevantes e permite reestruturar o conjunto de dados para instâncias por tempo.

Com o conjunto de dados organizado, a etapa de “Predição de Valores” apresenta técnicas e uso de ferramentas para realizar as predições por algoritmos de regressão linear múltipla, árvore de decisão, floresta aleatória e redes neurais.

Apuração de resultados para determinação do melhor modelo para cada ação.

A seguir será descrito aplicação da abordagem ForeSeries por meio de um estudo de caso com dados de séries temporais da Bovespa.

5.1 Estudo de Caso

O presente estudo de caso propõe aplicar a abordagem ForeSeries em bases de dados de séries temporais do mercado financeiro de ações, com objetivo de predizer os valores das ações no futuro. Assim, utilizou-se dados colhidos no dia 27/10/2017 e os valores futuros em 30/10/2017, para verificar qual método dessa abordagem conseguiu se aproximar dos valores reais. Em outras palavras, possibilitou-se verificar qual método “acertou mais”.

As bases de dados foram obtidas no site da B3 (Brasil, Bolsa, Balcão; 2017), a empresa que surgiu como controladora da Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&F Bovespa S.A.). A seguir são descritas as etapas da abordagem ForeSeries aplicadas ao estudo de caso demonstrado aqui.

5.1.1 Seleção das Fontes de Dados

A seleção das fontes de dados foi realizada no final do dia 27/10/2017, com os arquivos das cotações históricas obtidos no site da Bovespa (B3, 2017).

As bases de dados obtidas contêm cotações históricas de negociação desde de 1999 ao último dia fechado (27/10). Para facilitar a aplicação da abordagem ForeSeries somente as previsões de fechamento dos títulos que formam o índice Bovespa foram selecionadas, formada por 59 papéis listados na Tabela 1.1.

Tabela 1.1 - Relação dos códigos de negociação das ações do índice Bovespa em 27/10/2017.

ABEV3	CCRO3	ELET3	ITSA4	NATU3	SMLS3
BBAS3	CIEL3	ELET6	ITUB4	PCAR4	SUZB5
BBDC3	CMIG4	EMBR3	JBSS3	PETR3	TAEE11
BBDC4	CPFE3	ENBR3	KLBN11	PETR4	TIMP3
BBSE3	CPLE6	EQTL3	KROT3	QUAL3	UGPA3
BRAP4	CSAN3	ESTC3	LAME4	RADL3	USIM5
BRFS3	CSNA3	FIBR3	LREN3	RAIL3	VALE3
BRKM5	CYRE3	GGBR4	MRFG3	RENT3	VIVT4
BRML3	ECOR3	GOAU4	MRVE3	SANB11	WEGE3
BVMF3	EGIE3	HYPE3	MULT3	SBSP3	

Fonte: BM&F Bovespa S.A. (2017).

Como visto, a quantidade de ações selecionadas (59) permitiu a adequada validação da abordagem “ForeSeries”.

5.1.2 Transformação

A etapa de transformação permitiu o uso na linguagem R com o formato apropriado para a manipulação adequada em Linguagem R (CSV). A Figura 5.1 ilustra parte dos dados antes da transformação, na qual se observa o posicionamento fixo dos dados em formato texto.

Figura 5.1 – Fração do inicial do arquivo com as cotações históricas.

```
00COTAHIST.1999BOVESPA 20000103
011999010402ACES3       010ACESITA     ON *         R$  000000000005000000000000540000000
011999010402ACES4       010ACESITA     PN *         R$  000000000005900000000000640000000
011999010402ALPA3       010ALPARGATAS  ON *         R$  000000000550000000000055000000000
011999010402BAZA3       010AMAZONIA    ON *         R$  000000001500000000000160000000000
011999010402ANTA3       010ANTARCTICA  ON           R$  000000000260000000000026000000000
011999010402ARCZ6       010ARACRUZ     PNB          R$  000000000009400000000000950000000
011999010402AVPL3       010AVIPAL      ON *         R$  000000000016400000000001640000000
011999010402BESP3       010BANESPA     ON *EJ       R$  000000000385000000000038500000000
011999010402BESP4       010BANESPA     PN *EJ       R$  000000000510000000000051500000000
```

Fonte: BM&F Bovespa S.A. (2017).

Para a transformação foi desenvolvido um algoritmo (Anexo 1) em Java, o qual permitiu gerar os dados no formato CSV.

5.1.3 Unificação das Bases de Dados

A etapa de Unificação das Bases de Dados permitiu unificar as bases de dados em forma CSV. Para isso, foi necessário fazer o carregamento dos arquivos CSV conforme ilustrado na Figura 6.1, a qual apresenta o comando *read.csv*, para leitura das bases de dados e, *rbind*, para unificação.

Figura 6.1 – Leitura e junção dos dados por CSV.

```
cotacao_1999 <- read.csv("Dados/BovespaAnual/COTAHIST_A1999.csv")

cotacao_2000 <- read.csv("Dados/BovespaAnual/COTAHIST_A2000.csv")
cotacao_geral <- rbind(cotacao_1999, cotacao_2000)
```

Fonte: elaborado pelos autores.

Como observado a função *rbind* recebe dois conjuntos de dados (cotacao_1999 e cotacao_2000), que são os conjuntos de dados já carregados e, retorna um único conjunto de dados (cotacao_geral).

5.1.4 Limpeza e Organização dos Dados

A etapa de Limpar e Organizar dos Dados permitiu realizar a limpeza e organização dos dados, os dados originalmente estavam dispostos como na Tabela 2.1, com data, código de valor de fechamento.

Tabela 2.1 – Visualização das primeiras seis colunas da fonte de dados informada.

	DATA	CODNEG	PREULT
1	19990104	BBDC3	6,30
2	19990104	BBDC4	6,85
3	19990104	BBAS3	7,23
4	19990104	CMIG4	21,50
5	19990104	CPLE6	8,61
6	19990104	ELET3	20,30

Fonte: elaborado pelos autores.

Com a aplicação do código (Anexo 2) foi gerada uma organização adequada, como apresenta a Tabela 2.2, o resultado da organização dos dados com as datas orientação para as instâncias e as ações variando ao longo dos registros de forma cronológica.

Tabela 2.2 – Visualização de um recorte da base de dados ajustada.

	BBAS3	BBDC3	BBDC4	CMIG4	CPLE6	...
19990104	7,23	6,30	6,85	21,50	8,61	...
19990105	0,00	6,30	6,65	22,00	8,70	...
19990106	6,90	6,15	6,65	23,69	8,90	...
...	...	....	...	...	...	...

Fonte: elaborado pelos autores.

Para facilitar as manipulações nas próximas etapas, foi replicado os campos de uma instância para a seguinte, adicionando campos extras nos finais das colunas.

No fim desta etapa o conjunto de dados está estruturado e organizado para possibilitar predições, conforme descrito a seguir.

5.1.5 Predição de Valores

Essa etapa permite a execução sequencial dos métodos que predirão os valores futuros das ações, sendo: regressão linear múltipla, árvore de decisão, floresta aleatória e redes neurais, os quais serão descritos a seguir.

5.1.6 Regressão Linear Múltipla

Para utilizar a regressão linear múltipla utilizou-se um grau de correção superior a 0,95, considerando-se as 59 ações. Esse grau foi escolhido uma vez que se buscou valores muito correlatos, ou seja, próximo de um. Ressalta-se que o objetivo é determinar quais ações possuem uma aproximação em relação ao desempenho de flutuação do valor de mercado. Por exemplo, busca-se determinar quais ações possuem a mesma tendência se comparada às ações do Bradesco.

A Figura 8.1 ilustra o resultado pela busca por um conjunto de ações que possuem o grau de correlação superior a 0,95 (código disponível no Anexo 3), para a aplicação da regressão linear múltipla com variáveis selecionadas.

Figura 8.1 – Resultado do algoritmo de busca de pelo conjunto de ações com coeficiente de correlação superior a 0,95.

```
[1] "Ação Alvo:  QUAL3"
     [,1]
[1,] "TIMP3a"
[1] "Ação Alvo:  TIMP3"
     [,1]     [,2]
[1,] "QUAL3a" "VIVT4a"
[1] "Ação Alvo:  VIVT4"
     [,1]
[1,] "TIMP3a"
[1] "Ação Alvo:  ABEV3"
     [,1]
[1,] "BBSE3a"
[1] "Ação Alvo:  BBSE3"
     [,1]
[1,] "ABEV3a"
```

Fonte: elaborado pelos autores pelo R Studio.

A figura acima apresenta as relações da ação TIMP3, com a VIVT4 e QUAL3, que são respectivamente a Tim Participações e Telefônica Brasil ambas do setor de telecomunicações e a Qualicorp, uma gerenciadora de planos de saúde. Com essas informações foi montada a fórmula para a predição do valor das ações representado na Figura 8.2, a inclusão do valor da própria ação é imperativa, pois a mesma tem um alto valor de correlação.

Figura 8.2 - Geração do modelo e predição com a regressão linear múltipla.

```
modelo_lm <- lm(preult$TIMP3 ~
              preult$TIMP3a + preult$QUAL3a + preult$VIVT4a,
              data=preult)
preult$TIMP3p <- predict(modelo_lm, data=preult)
```

Fonte: elaborado pelos autores.

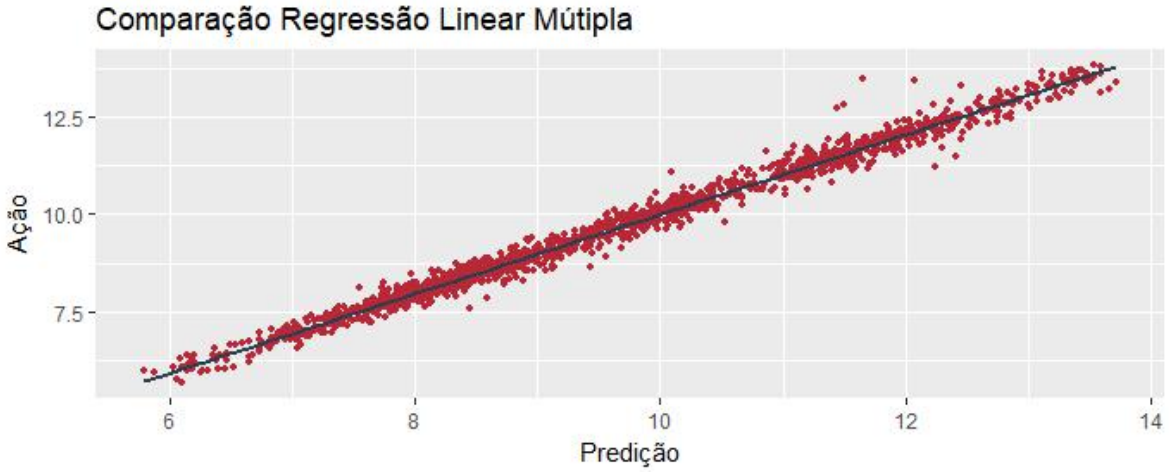
Neste estudo demonstra alguns resultados de predições com gráficos, com o código da Figura 8.3 foi gerado a Figura 8.4, que representa a comparação entre os valores reais das ações e os valores preditos.

Figura 8.3 - Plotagem do resultado da predição por regressão linear múltipla.

```
ggplot(preult, aes(x=TIMP3p, y=TIMP3)) +
  labs(title = "Comparação Regressão Linear Múltipla ",
           x = "Predição", y = "Ação") +
  geom_point(color='#B82837', size = 1) +
  geom_smooth(method=lm, se=FALSE, fullrange=FALSE, color='#2C3E50')
```

Fonte: elaborado pelos autores.

Figura 8.4 - Comparação dos resultados de predição por regressão linear múltipla e valores reais, a linha representa o objetivo desejado.



Fonte: elaborado pelos autores pelo R Studio.

Se aplicado a formula e dados adequados a regressão linear múltipla pode apresentar ótimos resultados.

5.1.7 Árvore de Decisão

Os conjuntos de dados de treino e teste foram gerados com o uso da função *sample* separando de forma aleatória 70% para treinamento e 30% para os testes de validação dos resultados do treinamento, Figura 9.1, o uso da do comando *set.seed* criou uma aleatoriedade controlada garantido a replicabilidade do estudo.

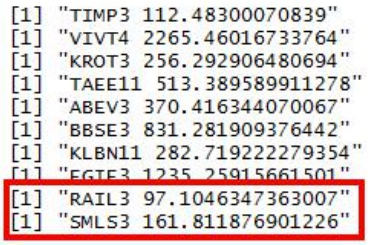
Figura 9.1 – Separação da amostra para teste e treino da base, com a função de *sample*, que "aleatoriamente" separa sorteia as linhas.

```
set.seed(157)
dfSample <- sample(1:nrow(preult),
                   size=nrow(preult)*0.7)
dfTreino <- preult[dfSample,]
dfTeste <- preult[-dfSample,]
```

Fonte: elaborado pelos autores.

Após a aplicação do código de busca dos erros médios (Anexo 4) foi observado, Figura 9.2, que a ação RAIL3 possui o valor mais baixo por isso foi escolhida como exemplo.

Figura 9.2 – Impressão dos resultados da busca pelos erros médios.



Fonte: elaborado pelos autores pelo R Studio.

O conjunto de dados passou por uma limpeza para remover os períodos em que as ações RAIL3 ainda não eram negociadas, com valor zerado pelo código na Figura 9.3.

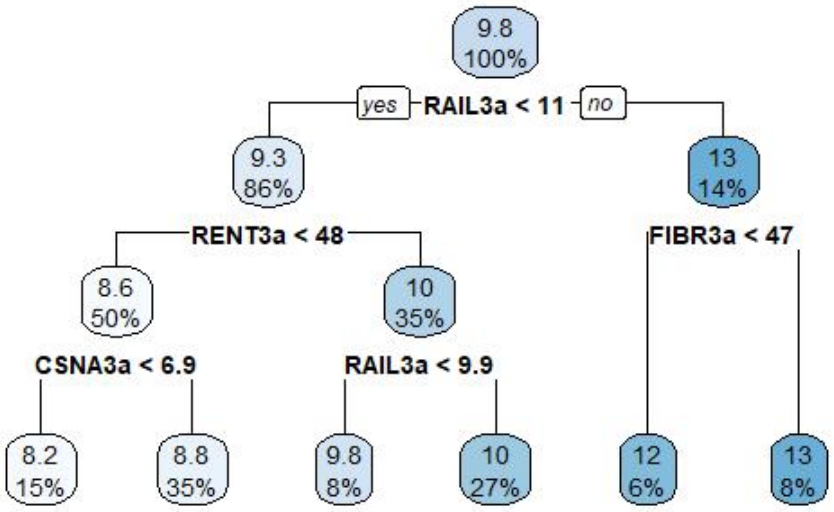
Figura 9.3 - Remover as linhas com a coluna RAIL3 zerados.

```
preult <- preult[(preult$RAIL3!=0),]
```

Fonte: elaborado pelos autores.

A Figura 9.4, gerada por um *rpart.plot* do modelo, demonstra a estrutura da árvore de decisão com os caminhos utilizados na busca dos resultados, atente-se as orientações de fluxo com um questionamento lógico.

Figura 9.4 – Estrutura da árvore de decisão utilizada na

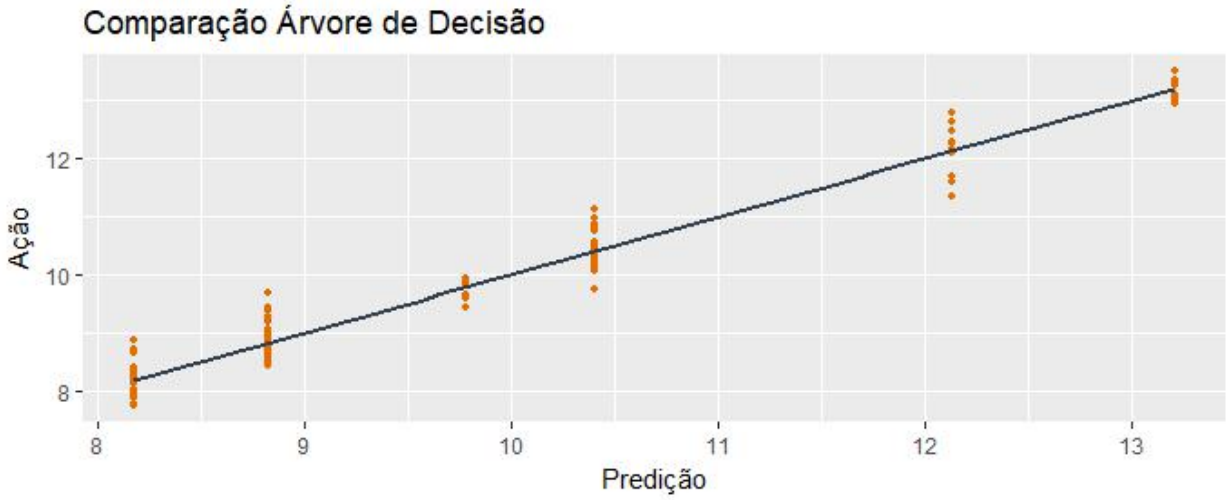


predição.

Fonte: elaborado pelos autores pelo R Studio.

Como observado na figura acima, os caminhos desenvolvidos na árvore de decisão somente geram seis resultados possíveis, o que também reflete nas previsões, observado na comparação da predição e realidade na Figura 9.5.

Figura 9.5 - Comparação dos resultados de predição por árvore de decisão e valores reais, a linha representa o objetivo desejado.



Fonte: elaborado pelos autores pelo R Studio.

A árvore de decisão não possui uma capacidade ampla para lidar com campos com grandes flutuações sendo recomenda para predição de resultados com menos variações no decorrer das instâncias, por exemplo temperatura média de meses do ano.

5.1.8 Floresta Aleatória

Com os conjuntos de dados de treino e teste gerados, com o código na Figura 9.1 na etapa anterior, foi gerado o modelo de floresta aleatória e a predição para uma determinada ação, como apresentado na Figura 10.1.

Figura 10.1 - A geração do modelo e predição para floresta aleatória.

```
modelo_fa <- randomForest(dfTreino[,i] ~ .,
                          data=dfTreino[,c(60:118)])
predicao <- predict(modelo_fa, dfTeste)
```

Fonte: elaborado pelos autores.

A ação RAIL3, participação sobre a América Latina Logística, com a aplicação de diversas árvores de decisão escolhendo as que apresentam o melhor incide de acerto, que pode ser observado pelo comando *getTree*, com a visualização na Figura 10.2.

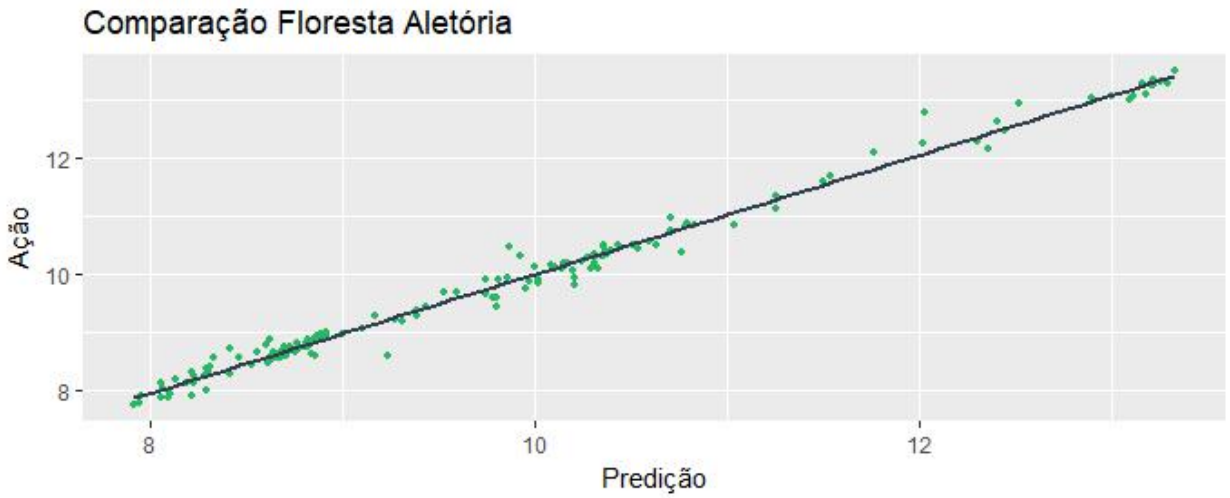
Figura 10.2 – Exemplo de valores de usados na construção da floresta aleatória.

	left daughter	right daughter	split var	split point	status	prediction
1	2	3	BRKM5a	42.055	-3	9.864324
2	4	5	RENT3a	46.235	-3	9.141379
3	6	7	ECOR3a	11.315	-3	12.485000
4	8	9	EQTL3a	53.825	-3	8.633214
5	10	11	RAIL3a	10.165	-3	10.059355
6	12	13	CMIG4a	8.485	-3	10.815000
7	14	15	GOAU4a	5.475	-3	13.041667
8	16	17	FIBR3a	37.325	-3	8.059000
9	18	19	RAIL3a	9.080	-3	8.758043
10	20	21	ABEV3a	18.715	-3	9.772105
11	22	23	KROT3a	15.125	-3	10.514167
12	0	0	<NA>	0.000	-1	11.590000
13	0	0	<NA>	0.000	-1	10.660000
14	0	0	<NA>	0.000	-1	12.216667
15	24	25	BBDC4a	36.155	-3	13.206667
16	0	0	<NA>	0.000	-1	8.254000
17	0	0	<NA>	0.000	-1	7.864000
18	26	27	VALE3a	29.800	-3	8.715238
19	0	0	<NA>	0.000	-1	9.207500
20	0	0	<NA>	0.000	-1	9.237500
21	28	29	ELET6a	18.400	-3	9.914667

Fonte: elaborado pelos autores pelo R Studio.

O gráfico na Figura 10.3 apresenta a comparação entre os valores reais e os preditos com um erro médio de 0,0117, para a ação RAIL3.

Figura 10.3 - Comparação dos resultados de predição por floresta aleatória e valores reais, a linha representa o objetivo desejado.



Fonte: elaborado pelos autores pelo R Studio.

A floresta aleatória sempre vai ser mais produtiva que uma árvore de decisão por ser uma composição de várias e seleção dos melhores resultados.

5.1.9 Rede Neural

Para a aplicação de Redes Neurais o conjunto de dados passou por uma transformação de escala, na qual os valores foram adequados em um intervalo [0 1]. Essa transformação de escala foi realizada utilizando a Equação (2).

$$v_{escala} = \frac{v_{original} - v_{mínimo}}{(v_{máximo} - v_{mínimo})} \quad (2)$$

Onde v original: valor de origem do conjunto de dados, v mínimo: valor mínimo encontrado na coluna, v máximo: valor máximo encontrado na coluna e v escala: valor no intervalo [0 1].

Assim, BBC3 em destaque da Figura 11.1 passa pela transformação pela aplicação do conceito (Anexo 5) que gerou o valor em destaque na Figura 11.2.

Figura 11.1 - Apresentação de um recorte do conjunto de dados.

	BBAS3	BBDC3	BBDC4	CMIG4	CPLE6	CSNA3	ELET3	ELET6
19990105	0.00	6.30	6.65	22.00	8.70	24.10	20.5	22.50
19990106	6.90	6.15	6.65	23.29	8.90	22.80	21.2	23.60
19990107	6.70	6.00	6.60	22.10	8.50	22.00	20.0	22.12
19990108	6.43	6.00	6.60	21.50	8.45	21.00	19.9	21.40
19990111	6.20	5.75	6.15	20.00	7.80	22.00	18.6	20.10
19990112	6.05	5.70	6.20	19.20	7.20	20.38	17.1	18.40

Fonte: elaborado pelos autores pelo R Studio.

Figura 11.2 - Apresentação de um recorte do conjunto de dados após a transformação para escalas.

	BBAS3	BBDC3	BBDC4	CMIG4	CPLE6	CSNA3	ELET3	ELET6
19990105	0.00000000	0.006655373	0.006319515	0.1614109	0.08015799	0.11536620	0.2900667	0.3198847
19990106	0.08298256	0.005747822	0.006319515	0.1731681	0.08480483	0.10914313	0.3026861	0.3410183
19990107	0.08057727	0.004840271	0.006066734	0.1623223	0.07551115	0.10531355	0.2810528	0.3125841
19990108	0.07733013	0.004840271	0.006066734	0.1568538	0.07434944	0.10052657	0.2792500	0.2987512
19990111	0.07456404	0.003327686	0.003791709	0.1431826	0.05924721	0.10531355	0.2558140	0.2737752
19990112	0.07276007	0.003025169	0.004044489	0.1358914	0.04530669	0.09755864	0.2287723	0.2411143

Fonte: elaborado pelos autores pelo R Studio.

Na geração do modelo de predição com o comando *neuralnet,* Figura 11.3, da biblioteca de mesmo nome, foi utilizado na camada oculta 12 neurônios, verdadeiro para neurônios externos da estrutura da rede o número máximo de passos usados para treinar o modelo, utilizou-se estes valores de forma arbitrária para garantir a execução de forma rápida da função para um melhor resultado e existe a possibilidade de aumentar essas disposições.

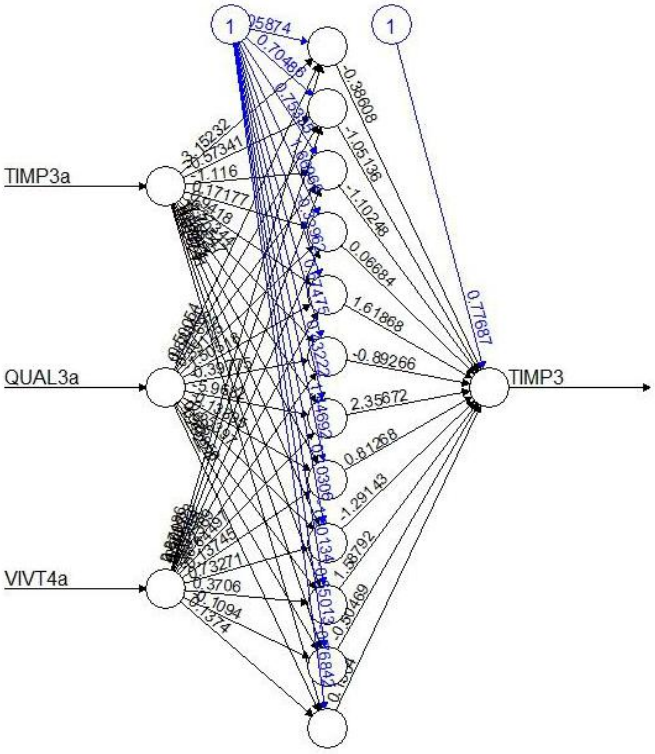
Figura 11.3 - Código para geração do modelo da rede neural com 12 neurônios na camada oculta, neurônios de fora da estrutura e 1000 com número máximo de passos.

```
modelo_rn = neuralnet(TIMP3 ~ TIMP3a + QUAL3a + VIVT4a,
            scaled, hidden = 12, linear.output = T, stepmax=1e4)
```

Fonte: elaborado pelos autores.

Com o comando *plot* no modelo observou o resultado (Figura 11.4) que apresenta a estrutura dos valores que formam a rede neural, vale ressaltar a presença de 12 neurônios na camada oculta e azul os neurônios externos.

Figura 11.4 – Visualização da estrutura da rede neural



Fonte: elaborado pelos autores pelo R Studio.

Para realizar a predição com o modelo definido, utilizou-se o comando *compute* com o as colunas selecionadas que retorna todos os valores utilizados na estrutura dos neurônios por isso utiliza-se somente a colunas *net.result*, com o resultado da rede demonstrado na Figura 11.5.

Figura 11.5 - Código de geração da predição com base das escalas e modelo gerados previamente e separação dos resultados.

```
predicao <- neuralnet::compute(modelo_rn,
                               scaled[,c('TIMP3a','QUAL3a','VIVT4a')])
preult$TIMP3_rn <- predicao$net.result
```

Fonte: elaborado pelos autores.

Após a predição, os valores preditos convertidos para a escala original conforme se observa na Figura 11.6.

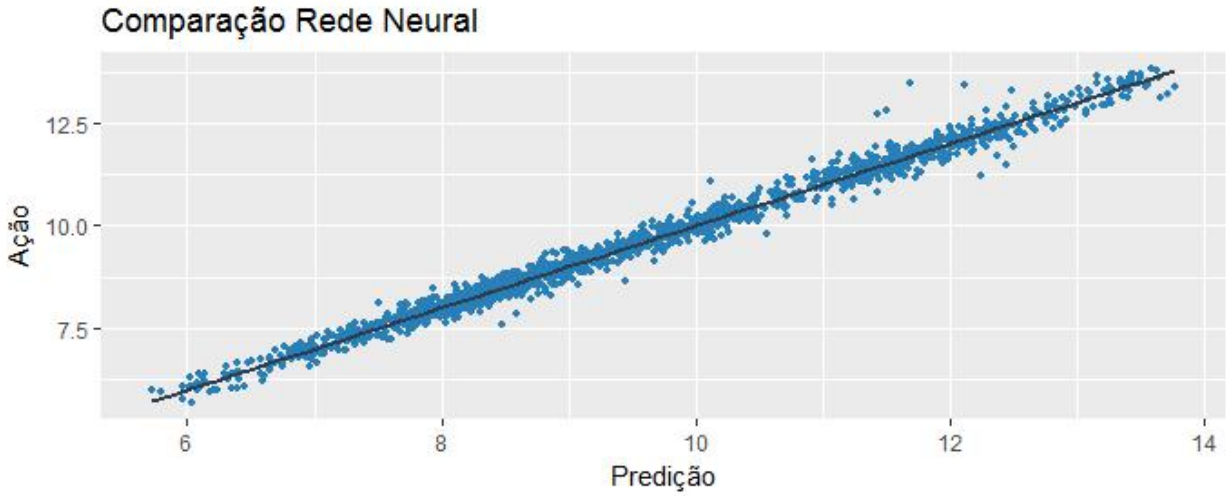
Figura 11.6 - Conversão da base de escalas para as prévias dos valores definidos com o uso das variáveis de características da escala.

```
preult$TIMP3_rn <- preult$TIMP3_rn
                   * (maxs[indice] - mins[indice]) + mins[indice]
```

Fonte: elaborado pelos autores.

A comparação entre os valores preditos já na escala original e os valores reais (que aconteceram de fato) foi visualizada conforme a Figura 11.7.

Figura 11.7 - Comparação dos resultados de predição por rede neural e valores reais, a linha representa o objetivo desejado.



Fonte: elaborado pelos autores pelo R Studio.

Como visto, a abordagem possibilitou a aplicação de quatro métodos diferentes para efetuar a predição dos valores futuros do mercado de ações e, ainda, propiciou a comparação dos valores preditos com os valores reais obtidos na Bovespa. A seguir são apresentados os resultados alcançados.

5.2 Resultados Alcançados

Para avaliar as predições efetuadas pelos quatros métodos previstos na abordagem ForeSeries, foi realizada uma comparação desses métodos com base em duas métricas: média do quadrado das diferenças entre os valores preditos e os reais (erro médio) e a diferença para o valor real e a predição no dia 30/10. A Tabela 3.1 apresenta os valores obtidos para a métricas para a cotação da América Latina Logística (RAIL3) no dia 30/10/2017. Ressalta-se que neste dia o valor real dessa cotação foi de R$ 12,80. Observa-se que o melhor resultado médio foi a floresta aleatória.

Tabela 3.1 - Resultados das predições para RAIL3 no dia 30/10/2017.

RAIL3	Ação no dia 30/10/2017: **R$ 12,80**		
Algoritmo	**Erro Médio**	**Predição**	**Diferença**
Regressão Linear Múltipla	**0,3083**	R$ 12,72	**-0,08**
Árvore de Decisão	0,0783	R$ 13,20	**+0,40**
Floresta Aleatória	**0,0117**	R$ 12,98	+0,18
Rede Neural	0,0503	R$ 12,99	+0,19

Fonte: elaborado pelos autores.

Nas mesmas condições temos que para a Tim Participações (TIMP3), que possuem um histórico de crescimento em ondas de seis meses, com os pontos baixos em junho e dezembro, com fechamento em 11.97, com uma queda acentuada após 4 dias de subidas, repetindo o melhor resultado anterior da floresta aleatória.

Tabela 3.2 - Resultados das predições para TIMP3 no dia 30/10/2017.

TIMP3	Ação no dia 30/10/2017: **R$ 11,97**		
Algoritmo	**Erro Médio**	**Predição**	**Diferença**
Regressão Linear Múltipla	0,0935	R$ 12,35	**+0,38**
Árvore de Decisão	**0,1090**	R$ 11,94	**-0,03**
Floresta Aleatória	**0,0068**	R$ 12,30	+0,33
Rede Neural	0,0433	R$ 12,34	+0,37

Fonte: elaborado pelos autores.

Segue as Tabelas 3.3, Tabelas 3.4 e Tabelas 3.5 apresentam os detalhes das previsões para ABEV3, BBSE3 e ECOR3 respectivamente, repetindo os resultados anteriores onde os erros médios.

Tabela 3.3 - Resultados das predições para ABEV3 (AmBev) no dia 30/10/2017.

ABEV3	Ação no dia 30/10/2017: **R$ 20,96**		
Algoritmo	**Erro Médio**	**Predição**	**Diferença**
Regressão Linear Múltipla	**0,3143**	R$ 20,51	**-0,45**
Árvore de Decisão	0,1102	R$ 21,32	+0,36
Floresta Aleatória	**0,0090**	R$ 21,00	**+0,04**
Rede Neural	0,0552	R$ 20,90	-0,06

Fonte: elaborado pelos autores.

Tabela 3.4 - Resultados das predições para BBSE3 (BB Seguridade Participações SA) no dia 30/10/2017.

BBSE3	Ação no dia 30/10/2017: **R$ 28,00**		
Algoritmo	**Erro Médio**	**Predição**	**Diferença**
Regressão Linear Múltipla	0,5532	R$ 28,74	**+0,74**
Árvore de Decisão	**0,8487**	R$ 27,81	**-0,19**
Floresta Aleatória	**0,0512**	R$ 28,65	+0,65
Rede Neural	0,3320	R$ 28,32	+0,32

Fonte: elaborado pelos autores.

Tabela 3.5 - Resultados das predições para ECOR3 (EcoRodovias) no dia 30/10/2017 com suas margens de erro por cada algoritmo aplicado.

ECOR3	Ação no dia 30/10/2017: **R$ 12,16**		
Algoritmo	**Erro Médio**	**Predição**	**Diferença**
Regressão Linear Múltipla	0,0975	R$ 12,36	+0,20
Árvore de Decisão	**0,3776**	R$ 13,31	**+1,15**
Floresta Aleatória	**0,0077**	R$ 12,30	+0,14
Rede Neural	0,0525	R$ 12,26	**+0,10**

Fonte: elaborado pelos autores.

O algoritmo que apresentou o melhor resultado em todos os casos foi a floresta aleatória seguida pela rede neural.

Existem correções para ser realizadas nas buscas, como reduzir o escopo para um período de maior controle, pois muitas ações embora sejam significativas iniciaram sua negociação somente em um período muito recente, algumas tiveram *split* e *merge*, separação e

junção de pacotes de ações respectivamente, no período que pode influenciar o resultado dos dados.

Outras técnicas de predição também podem se aplicar para séries temporais como *forecast, bootstrap, naive bayes* entre outras, mas não foram tratadas aqui por falta de tempo.

A ampliação das informações provenientes de fontes de dados abertos como taxas de crédito do mercado que facilitam ou dificultam a busca pelo investimento em bolsas de valores, índices de comércio, produção agrícola, extração mineral é liberado a cada início de trimestre, assim como a aplicação de outros algoritmos de regressão ou aprendizagem de máquina, como redes neurais, juntamente com algum método para interpretação de dados não numéricos como o sistema de escores para notícias relacionadas ao mercado de financeiro.

Com a análise dos resultados e a necessidade de refino e ampliação dos dados e métodos aplicados podemos concluir que a predição é possível e que pode levar a resultados cada vez mais próximos dos reais sendo necessários mais estudos para se encontrar uma metodologia que alcance valores mais próximos dos reais.

6. Conclusão

O presente projeto permitiu o desenvolvimento da abordagem ForeSeries, a qual possui quatro métodos de predição de valores em séries temporais: regressão linear múltipla, floresta aleatória e redes neurais. Para validação da abordagem foi utilizando um estudo de caso de ações da Bovespa, no qual objetivou-se comparar os valores de 59 ações em dois momentos distintos: 27/10/2017 e 30/10/2017. Dessa forma, foi possível verificar qual predição da abordagem ForeSeries proporcionou melhores “acertos”.

Como resultado futuro, sugere-se a ampliação dos modelos de predição e dados pertinentes ao ambiente analisado.

Por fim, ressalta-se que a abordagem ForeSeries é capaz de predizer valores em séries temporais com uma margem considerável de acertos.

O desenvolvimento do método apresentou resultados próximos dos esperados mostrando ser um ótimo início de desenvolvimento de técnicas mais apuradas e automatizadas, também demonstrou que mesmo em um ambiente volátil como a negociação de participações de empresas a predição de valores em séries temporais é possível.

Referências

ALCÂNTARA, Igor. Workshops ministrados online por Igor Alcântara com os temas **Fundamentos da Ciência de Dados**, **Linguagem R** e **Introdução à Machine Learning** nos dias 13 e 27 de agosto e 24 de setembro de 2017;

ASSAF Neto, Alexandre. **Mercado Financeiro**. 12ª edição São Paulo: Editora Atlas S.A., 2014;

B3. **Cotações Históricas** Anuais BM&FBOVESPA, de 04 de janeiro de 1999 a 27 de outubro de 2017. Disponível em: <http://www.bmfbovespa.com.br/pt_br/servicos/market-data/historico/mercado-a-vista/cotacoes-historicas/>. Acesso em: 27.out.2017;

B3. **Layout do Arquivo - Cotações Históricas** de 13 de abril de 2017. Disponível em: <http://www.bmfbovespa.com.br/lumis/portal/file/fileDownload.jsp?fileId=8A828D294E9C618F014EB7924B803F8B>. Acesso em: 11.ago.2017;

BORGES JUNIOR, Sergio Ricardo et al. **SEnsembles** - uma abordagem para melhorar a qualidade das correspondências de instâncias disjuntas em estudos observacionais explorando características idênticas e ensembles de regressores, 2016;

CARVALHO, Hialo Muniz. **Aprendizado de máquina voltado para mineração de dados**: árvores de decisão, 2015;

COELHO, Antônio C.; CUNHA, Jacqueline VA. Regressão linear múltipla. **Análise multivariada: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas**, p. 131-231, 2007;

DUHIGG, Charles. **O poder do hábito**: Por que fazemos o que fazemos na vida e nos negócios. 1ª edição Rio de Janeiro: Objetiva, p. 195-225, 2012;

FRITSCH, Stefan et al. **Training of Neural Networks**. 2016. Disponível em: https://cran.r-project.org /web/packages/neuralnet/neuralnet.pdf. Acesso em: 27.nov.2017;

HENRIQUES, Daniela Aparecida; DA COSTA, Helder Rodrigues. Big Data: como utilizar a extraordinária quantidade de informações coletadas por novas tecnologias para obter vantagens competitivas. **Revista Pensar**, v. 3, n. 1, 2014;

GUIMARÃES, Paulo Ricardo B. **Análise de Correlação e medidas de associação**. 2013. Disponível em: <https://docs.ufpr.br/~jomarc/correlacao.pdf>. Acesso em: 01.set.2017;

PROVOST, Foster, FAWCETT, Tom. **Data Science para Negócios**. Rio de Janeiro, RJ: Alta Books, 2016;

LIAW, Andy; WIENER, Matthew. **Breiman and Cutler's Random Forests for Classification and Regression**. 2015. Disponível em: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Acesso em: 27.nov.2017;

MAIA JÚNIOR, Humberto. **Da terra brotam os dados**, Revista Exame. Publicada em 02.out.2014. Disponível em: <https://exame.abril.com.br/revista-exame/da-terra-brotam-os-dados/. Acesso em: 29.set.2017;

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. **Big Data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro, RJ: Elsevier, 2013;

R FOUNDATION, The. 2017. **The R Project for Statistical Computing**. Disponível em: <https://www.r-project.org/>. Acesso em: 15.ago.2017;

RAGSDALE, Cliff T. **Modelagem de planilha e análise de decisão**; Tradução da 7ª edição norte-americana, 1ª Edição Editora Cengage, 2014;

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados**: Com aplicações em R; 1ª Edição Rio de Janeiro Editora Elsevier, 2016;

THERNEAU, Terry; ATKINSON, Beth; RIPLEY, Brian. **Recursive Partitioning and Regression Trees**. 2017. Disponível em: https://cran.r-project.org/web/packages/rpart/rpart.pdf. Acesso em: 27.nov.2017;

TRIOLA, Mario F. **Introdução à Estatística**: Atualização da Tecnologia, 11ª Edição Editora LTC, 2013.

Anexos

Anexo 1 - Algoritmo em Java para converter os arquivos em texto com posições pré definidas para csv somente com os campos desejados.

```
import java.util.*;

public class BSPtoCSV {
      public static void main(String[] args) {
            Scanner sc = new Scanner(System.in);
            String leitura, codneg, data, preult, codnegCompare;
            System.out.println("DATA,CODNEG,PREULT");
            while (sc.hasNext()) {
                  leitura = sc.nextLine();
                  if ("01".equals(leitura.substring(0, 2))) {
                        data = leitura.substring(2, 10).trim();
                        codneg = leitura.substring(12, 24).trim();
                        codnegCompare = "|" + codneg + "|";
                        preult = leitura.substring(108, 119).trim()
                         + "." + leitura.substring(119, 121).trim();

                        if
("|ABEV3|BBAS3|BBDC3|BBDC4|BBSE3|BRAP4|BRFS3|BRKM5|BRML3|BVMF3|CCRO3|
CIEL3|CMIG4|CPFE3|CPLE6|CSAN3|CSNA3|CYRE3|ECOR3|EGIE3|ELET3|ELET6|EMB
R3|ENBR3|EQTL3|ESTC3|FIBR3|GGBR4|GOAU4|HYPE3|ITSA4|ITUB4|JBSS3|KLBN11
|KROT3|LAME4|LREN3|MRFG3|MRVE3|MULT3|NATU3|PCAR4|PETR3|PETR4|QUAL3|RA
DL3|RAIL3|RENT3|SANB11|SBSP3|SMLS3|SUZB5|TAEE11|TIMP3|UGPA3|USIM5|VAL
E3|VIVT4|WEGE3|".contains(codnegCompare)) {
                              System.out.println(
                                    data + "," + codneg + "," + preult);
                        }
                  }
            }
      }
}
```

Fonte: elaborado pelos autores.

Anexo 2 - Organizar a estrutura de colunas e linhas da base de dados.

```
# Alterar a estrutura de entre data e código de negociação
preult <- reshape::cast(cotacao, DATA ~ CODNEG,
                          value="PREULT", fun.aggregate="sum")

# Renomear campos
row.names(preult) <- preult[,1]
preult <- preult[,-1]
```

Fonte: elaborado pelos autores.

Anexo 3 - A busca por variáveis que apresentem um grau de correlação maior que 0,95, imprimindo os resultados para as tendências desejadas.

```
for(i in 1:59) {
  relacoes <- c()
  acoes <- c()
  for(j in 60:ncol(preult)) {
      if (substr(colnames(preult)[i], 1, 4)
          != substr(colnames(preult)[j], 1, 4)) {

           correlacao <- cor(preult[,i], preult[,j],
                             use="complete.obs", method="pearson")
           if (abs(correlacao) > 0.95) {
              relacoes <- cbind(relacoes, preult[,j])
              acoes <- cbind(acoes, colnames(preult)[j])
           }
      }
  }

  if (!is.null(relacoes)) {
    print(paste("Ação Alvo: ", colnames(preult)[i]))
    print(acoes)
  }
}
```

Fonte: elaborado pelos autores.

Anexo 4 - Iterar as ações com os valores anteriores e imprimir o erro médio.

```
# Montagem de texto com os campos dos valores anteriores
colunas <- paste(colnames(dfTeste[,60:118]),
                 collapse = "+")

# geração de modelo, aplicação da predição e impressão do erro
médio
for(i in 1:59) {
  modelo_ad <- rpart(paste("dfTreino[,i] ~", colunas),
                     data=dfTreino)
  predicao <- predict(modelo_ad, dfTeste)

  erro <- mean((predicao - dfTeste[,i])^2, na.rm = TRUE)
  print(paste(colnames(dfTeste)[i], erro))
}
```

Fonte: elaborado pelos autores.

Anexo 5 - Converter os valores para escala e para um conjunto de dados.

```
# obtém os valores máximo e mínimo das colunas.
maxs <- apply(preult, 2, max)
mins <- apply(preult, 2, min)

# efetuar a transformação para a mesma escala.
scaled <- as.data.frame(as.matrix(
    scale(preult, center = mins, scale = maxs - mins)))

# renomeia as colunas obtidas.
rownames(scaled) <- rownames(preult)
colnames(scaled) <- colnames(preult)
```

Fonte: elaborado pelos autores.