

ARTIGO ORIGINAL

## Modelagem, implementação e avaliação de estratégias de negociação baseadas em algoritmos de aprendizado de máquina para o mercado financeiro

### Modeling, implementation and evaluation of negotiation strategies based on learning algorithms of machine for the financial market

Eduardo Jabbur Machado<sup>1</sup>, Carlos Alberto Silva de Assis<sup>1</sup>, Adriano César Machado Pereira<sup>1,2</sup>

<sup>1</sup>Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), <sup>2</sup>Universidade Federal de Minas Gerais (UFMG)

\*ejabbur@gmail.com; carlos.assis.79@gmail.com; adrianoc@gmail.com

Recebido: 13/02/2019. Revisado: 08/01/2020. Aceito: 23/03/2020.

#### Resumo

O investimento no mercado de ações é uma das formas mais rápidas e atrativas de obter lucros consideráveis em um curto espaço de tempo. Porém, devido a grandes variações e oscilações neste tipo de mercado, os investidores estão sujeitos à riscos que podem acarretar também em grandes prejuízos. Para evitar que outros estudantes e interessados da área de mercado financeiro tenham que despender muito tempo de suas pesquisas na implementação de algoritmos e possam dedicar esforços em criar, validar e aprimorar suas estratégias de negociação, propõe-se neste trabalho o projeto e implementação de um arcabouço automatizado constituído de 5 etapas: Extração de Dados, Caracterização e Transformação de Dados, Classificação de Modelos de Previsão de Tendências, Estratégia de Operação e a Análise de Resultados. Durante as simulações, avaliou-se dados de cotações históricas de 9 ativos negociados na Bolsa Brasil Balcão (B3), para um período de 741 na etapa de Validação, para os 8 modelos de previsão de tendências propostos. Como resultados e validação do arcabouço proposto, serão apresentados para cada um dos ativos e modelos de previsão de tendência, uma tabela consolidada contendo dados (desempenho, operação / risco e estatísticos) e 2 gráficos: série de preço de fechamento e série de evolução de capital acumulado (retornos líquido e bruto e custo operacional).

**Palavras-Chave:** Algoritmos de aprendizado de máquina; Caracterização e transformação de dados; Estratégias de negociação; Medidas de risco; Mercado de ações.

#### Abstract

Investing in the stock market is one of the fastest and most attractive ways to make considerable profits in a short period of time. However, due to large variations and fluctuations in this type of market, investors are subject to risks that can also result in large losses. In order to avoid that other students and interested in the financial market area have to spend a lot of time on their research in the implementation of algorithms and can dedicate efforts in creating, validating and improving their trading strategies, this work proposes the design and implementation of a automated framework consisting of 5 stages: Data Extraction, Data Characterization and Transformation, Classification of Trend Forecasting Models, Operation Strategy and Results Analysis. During the simulations, historical quotation data of 9 assets traded on the Brazil Balcão Exchange (B3) was evaluated, for a period of 741 in the Validation stage, for the 8 proposed trend forecasting models. As a result and validation of the proposed framework, a consolidated table containing data (performance, operation / risk and statistics) and 2 graphs: series of closing price and series of accumulated capital (liquid and gross returns and operating cost) evolution of trends will be presented for each of the assets and trend forecasting models.

**Keywords:** Data characterization and transformation; Machine Learning Algorithms; Negotiation strategies; Risk Measures; Stock Market.

## 1 Introdução

O mercado financeiro é constituído de agentes investidores que compram e vendem ativos com o objetivo de obter lucros. Os agentes investidores estão sujeitos a alguns fatores, como: impulso, medos e vontade de ganhar, que podem interferir em suas decisões. A fim de minimizar a interferência desses fatores em suas decisões, os agentes investidores se baseiam em vários fundamentos de mercado como a análise técnica que será abordada neste trabalho.

A análise técnica surgida no século XX através das teorias propostas por *Charles Dow* ([Hamilton, 1922](#)), as quais identificam o comportamento e os movimentos futuros de ativos baseado em estatísticas do histórico de preço, volume e indicadores, com o estudo de dados numéricos relativos e dos gráficos das empresas que mostravam o comportamento das ações ao longo do tempo.

O problema de pesquisa abordado neste artigo é um tema interdisciplinar que envolve diversas áreas do conhecimento, como por exemplo, matemática, estatística, economia, ciência de dados e inteligência computacional, de forma integrada e aplicada. Diante deste cenário, destaca-se 3 questões a serem investigadas:

- Tratamento dos Dados: Qual a melhor forma e/ou técnicas de ciências dos dados para selecionar, preparar e analisar os dados de séries de cotação histórica sobre o mercado financeiro?
- Modelo de Previsão: Como configurar, treinar, medir, combinar e selecionar um ou vários algoritmos de classificação para a geração de previsão de tendências que garanta uma taxa mínima de aprendizado em relação à sua precisão como medida de desempenho?
- Modelo de Operação: De que forma pode-se aplicar uma série de previsão de tendências como estratégia de operação, buscando resultados cujo percentual de retorno financeiro seja viável, rentável e que supere os valores de referência dos *baselines* para a aplicação em um ambiente real do mercado financeiro?

A ideia deste estudo é avaliar a viabilidade e o grau de risco de investimento das propostas de estratégias de operação utilizando algoritmos de aprendizado de máquina combinados e a possibilidade de aplicação futura destas estratégias de operação em um cenário real no mercado financeiro.

Desta forma, propõe-se uma metodologia composta por 5 etapas (coleta, transformação, classificação, operação e análise de resultados), que utilizam dados históricos de cotações de ações, preço e volume, disponibilizados pela Brasil Bolsa Balcão (B3) de 9 códigos de ativos (i.e, BBAS3, JBSS3, KROT3, LAME4, MRVE3, NATU3, PETR4, RADL3 e TIMPL3) de segmentos distintos de mercado, como entrada do modelo de previsão de tendência utilizando algoritmos de aprendizado de máquina. Tais previsões geram estratégias através de um arcabouço que analisa diversos dados tais como retorno financeiro, estatísticas de desempenho, custo operacional e risco.

O restante deste artigo está organizado da seguinte

forma: a [Seção 2](#) apresenta os trabalhos relacionados selecionados através de uma metodologia sistemática de revisão de bibliográfica. A apresentação de alguns conceitos sobre custos operacionais, algoritmos de aprendizado de máquina, métricas de avaliação de desempenho e risco serão detalhadas na [Seção 3](#). A metodologia proposta composta de 5 etapas (Coletar, Transformar, Classificar, Operar e Analisar) está descrita na [Seção 4](#). A [Seção 5](#) apresenta a simulação realizada juntamente com a apresentação dos resultados. Finalmente, a [Seção 6](#) descreve as conclusões e direções futuras.

## 2 Trabalhos Relacionados

Com o propósito de identificar e extrair conceitos, fundamentos e trabalhos relacionados de uma forma metodológica, realizou-se uma revisão sistemática da literatura (RSL) adotando o conjunto de diretrizes propostas por [Kitchenham and Charters \(2007\)](#). Sua principal característica é realizar a definição de um protocolo de avaliação como pré-requisito antes de avaliar os trabalhos que serão selecionados.

A seguir serão apresentadas as principais pesquisas selecionadas pela RSL com o detalhamento de alguns estudos que abordam aplicações de aprendizado de máquina com mercado financeiro.

[White \(1988\)](#), [Niaki and Hoseinzade \(2013\)](#), [Kaupa \(2013\)](#) de [Sousa and Lopes \(2014\)](#), [Melo \(2012\)](#), [Oliveira \(2015\)](#) e [Iglesias Caride et al. \(2018\)](#) dedicaram esforços no sentido de desenvolver modelos de classificação de tendências de ativos financeiros utilizando na maioria das vezes algoritmos de aprendizado de máquina. Como fonte de dados, utilizaram dados financeiros provenientes de plataformas proprietárias (e.g. *Bloomberg*, *Thompson Reuters*), redes sociais, fóruns de discussão na Internet, dentre outras fontes livremente disponíveis.

Trabalhos desenvolvidos por [Wei and Cheng \(2012\)](#), [Hrasko et al. \(2015\)](#), [Shen et al. \(2015\)](#), [Liang et al. \(2017\)](#), [Chong et al. \(2017\)](#) e [Assis et al. \(2018\)](#) estenderam diferentes arquiteturas de redes neurais artificiais profundas conhecidas como *deep learning* para avaliar predição de séries temporais, como por exemplo, *Feedforward Neural Network (FNNs)*, *Continuous Restricted Boltzmann Machine (CRBM)*, *Restricted Boltzmann Machines (RBM)*, *Recurrent Neural Network (RNN)*, *Deep Belief Network (DBN)*, *Principal Component Analysis (PCA)*, *AutoEncoder (AE)* e *Support Vector Machine (SVM)*, com o objetivo de gerar um modelo híbrido realizando inicialmente a redução de dimensionalidade e seleção de características através do algoritmo de categoria não-supervisionado RBM para a previsão de retornos de preços futuros.

Através da RSL foram mapeados os principais trabalhos relacionados à caracterização, modelagem e previsão de séries temporais em específico do mercado financeiro, utilizando para isso, técnicas, recursos e modelos computacionais de aprendizado de máquina associados à indicadores financeiros e de tendências.

Para facilitar a análise e identificação das referências, criou-se a [Tabela 1](#) na qual é comparado o que foi realizado nos trabalhos coletados destacando os seguintes

Tabela 1: Comparação entre os trabalhos selecionados.

Referências e Ano	Fontes de Dados	Algoritmos e Modelos	Métricas de Avaliação
(Assis et al., 2018)	B3	RBM e SVM	Acurácia
(Liu, 2017)	Índice SSE e FTSE 100	GRBM e SVM	MSE, RMSE, MAPE
(Lee, 2017)	Dados corporativos	RBM + BP, DBN e FNN	RMSE.
(Barboza et al., 2017)	Dados corporativos	SVM e Random Forest	Precisão
(Nametala, 2017)	B3	RNA	Acurácia
(Dixon et al., 2016)	B3	DNN	Precisão
(Xiong, 2016)	Índice S&P500	LSTM e RIDGE	MAPE
(Jia, 2016)	Índice S&P500	LSTM, RIDGE e LASSO	Acurácia
(Booth, 2016)	Índices DAX e FTSE100	MLP, SVR e RF	Acurácia
(Ferreira, 2016)	StockTwits	-	Análise de correlação
(Santos, 2016)	B3 e Twitter	-	Análise de correlação
(Mansouri, 2016)	Índice S\&P100	LSTM	Precisão
(Hrasko et al., 2015)	CAT's benchmark data	MLP, DBN, RBM, ANN	MSE e RMSE
(Shen et al., 2015)	Dolar	DBN, CBRBM, FFNN	RMSE, MAE, MAPE
(Hrasko, 2015)	Dow Jones	GBRBM+FNN e BP	RMSE
(Liu, 2015)	Bloomberg e S\&P500	CNN	Acurácia e CC de Matthews
(Ding et al., 2015)	Índice S&P500	WB-NN e WB-CNN	ACUR, CMAT
(Cao et al., 2015)	Dow Jones	ARIMA, RL e ANN	ACUR, PREC
(Sharang, 2015)	US Treasury	PCA, DBN e RBM	Recall
(Estrada, 2015)	Índice S\&P500	RBM e DBN-MLP	Acurácia
(Aamodt, 2015)	S\&P500, NASDAQ	SVR, FFNN, CRBMe CNN	Acurácia
(Zhang, 2015)	Dow Jones	RNN, RBM, RF e SVM	Acurácia
(de Souza Roca and Mól, 2015)	B3	PCA, MLP e GARCH	MSE, RMSE, MAE
(da Silva, 2014)	B3	NB, RL e SVM	ACUR
(Ramalho, 2014)	B3	MLP + BP	MSE, RMSE e MAPE
(Vengertsev, 2014)	B3	SVR, MLP e RBM + MLP	Acurácia

pontos: referências, fontes de dados utilizada, algoritmos e ou modelos de previsão de tendência propostos e por último, as medidas de avaliação e desempenho (Métricas) apresentadas como resultados.

Em relação às fontes de dados utilizadas pelos trabalhos selecionados na Tabela 1, cerca de 70% utilizam dados de índices e indicadores de bolsas do mercado norte-americano. Esta incidência é justificada pela grande quantidade de fontes de dados disponíveis na Web e em plataformas proprietárias e, principalmente, pela estabilidade do mercado financeiro norte-americano em relação à outros mercados, em especial, o mercado financeiro brasileiro.

Os algoritmos mais utilizados de forma independente e ou combinado foram as Máquinas Restrita de Boltzmann (RBM) com 18% e, em seguida, o Perceptron de Múltiplas Camadas (MLP) e o Deep Belief Networks (DBN), ambos com 13%. Vários algoritmos de aprendizado de máquina e ou técnicas de análise de classificação que ocorreram com pouca frequência totalizaram a proporção de 31%. Visualizamos a possibilidade de utilização do algoritmo RBM de categoria não-supervisionada associado a outros algoritmos de classificação tradicionais de categoria supervisionada que serão detalhados a seguir na Seção 3 de fundamentação teórica.

Entre as métricas de avaliação empregadas nos algoritmos e os modelos de previsão de tendências, utilizados nos trabalhos selecionados, cerca de 90% utilizam a métrica de avaliação de medida de desempenho (Acurácia), acrescida de medidas de estatística de estimação de erro, como por exemplo, Erro Quadrado Médio (MSE), Raiz do Erro Quadrado Médio (RMSE) e Média Percentual Absoluta do Erro(MAPE).

Os autores selecionados na Tabela 1, totalizaram 23 trabalhos, sendo 11 artigos (Liu, 2017, Lee, 2017, Barboza et al., 2017, Dixon et al., 2016, Xiong, 2016, Jia, 2016, Hrasko, 2015, Liu, 2015, Ding et al., 2015, Cao et al., 2015, Sharang, 2015), 11 dissertações de mestrado (Nametala, 2017, Ferreira, 2016, Santos, 2016, Mansouri, 2016, Estrada, 2015, Aamodt, 2015, Zhang, 2015, de Souza Roca and Mól, 2015, da Silva, 2014, Ramalho, 2014, Vengertsev, 2014) e uma tese de doutorado (Booth, 2016) de forma convergente, realizam a utilização de indicadores financeiros, análise de informações da Web, técnicas de ciência dos dados (normalização e transformação, caracterização e análise) e aplicações de aprendizado de máquina.

Com o propósito de conhecer o estado da arte atual para propor direções e novos caminhos a serem investigados, verificou-se várias dificuldades em replicar os experimentos realizados nestes trabalhos relacionados, devido não existir um padrão de metodologias, arquiteturas, ferramentas, acesso às fontes de dados, dentre outros, comum aos trabalhos relacionados.

Devido ao tema desta pesquisa ser interdisciplinar, o principal desafio comum a todos estes trabalhos relacionados, investigados apresentados na Tabela 1, é tentar caracterizar, compreender e prever o comportamento dos preços e conseqüentemente a direção do fluxo das tendências dos ativos negociados no mercado de ações, independente da técnica utilizada.

Observou-se após o levantamento dos trabalhos relacionados, uma oportunidade e necessidade de construir um arcabouço automatizado para a modelagem e a avaliação de estratégias de investimento, utilizando para isso, conceitos de teorias econômicas, dados extraídos



da Web e algoritmos de aprendizado de máquina, com a finalidade de implementar modelos classificadores de previsão de tendências, que pudessem ser parametrizáveis e disponibilizado para a comunidade acadêmica.

Esta pesquisa diferencia-se dos demais trabalhos relacionados da Tabela 1 devido ao arcabouço proposto gerar de forma automatizada, através da Linguagem Python, gráficos e tabelas contendo os resultados estatísticos, desempenho (acurácia, precisão e retorno financeiro), risco, custos operacionais, dentre outros, para os 8 modelos classificadores de previsão de tendências que serão detalhados na Seção 4.

### 3 Fundamentação Teórica

A seguir serão apresentados alguns conceitos sobre custos operacionais, algoritmos de aprendizado de máquina utilizados como modelos de previsão de tendências e algumas métricas de avaliação de desempenho e risco.

#### 3.1 Custos Operacionais do Mercado de Ações

O investimento direto em ações e outros ativos de renda variável envolve alguns custos operacionais, tais como, taxa de corretagem, taxa de custódia, emolumentos e imposto de renda, que podem sofrer variações de valores conforme o tipo de transação realizada (B3-Variável, 2020), os quais serão detalhados a seguir:

- **Taxa de corretagem:** é um valor fixo ou percentual sobre o capital investido cobrado pelas corretoras para intermediar as operações de compra ou venda de ações junto à bolsa de valores;
- **Emolumentos e liquidação:** são valores cobrados pela B3 através de uma taxa fixa para cada tipo de operação, com a finalidade de cobrir custos operacionais que incidem na compra e venda direta de ações. As operações normais, compra e venda de ações efetuadas em pregões distintos, tem incidência total de 0,0325%. Já as operações de *daytrade*, compra e venda de um ativo em um mesmo dia, tem a incidência de 0,025%;
- **Taxa de custódia:** é um valor cobrado pelas corretoras para intermediar as operações de compra ou venda de ações junto à B3 e, dependendo da corretora o cliente pode até ficar isento desta taxa; e
- **Imposto de Renda:** o lucro obtido ao se operar renda variável, o chamado ganho líquido, está sujeito à cobrança de imposto de renda. Ganhos líquidos com operações comuns são tributados em 15%. Já aqueles obtidos com operações *daytrade* são tributados em 20%. Há isenção de IR para vendas no valor de até 20 mil reais em um único mês.

A apuração e o recolhimento do imposto de renda sobre o ganho líquido na negociação de ativos de renda variável são de responsabilidade do próprio investidor. Quando houver uma venda sujeita à tributação, o investidor deve emitir um Documento de Arrecadação de Receitas Federais (DARF) e pagá-lo até o último dia útil do mês seguinte ao da operação.

#### 3.2 Algoritmos de Aprendizado de Máquina

Aprendizado de máquina são métodos de análise de dados que automatizam o desenvolvimento de modelos analíticos, usando algoritmos que aprendem interativamente a partir de dados, permitindo que os computadores encontrem respostas e ou representações ocultas sem serem explicitamente programados para procurar algo específico, através de um programa de treinamento de recompensa e punição, sendo um ciclo semelhante a aprendizagem do cérebro humano (Chorafas, 1988).

Os modelos de aprendizado de máquina são subdivididos em quatro grupos de algoritmos de aprendizado: Supervisionado, Não Supervisionado, Semi-Supervisionado e Por Reforço.

- **Supervisionado:** são apresentadas ao computador exemplos de entradas e saídas desejadas, com objetivo de aprender uma regra geral que mapeia as entradas para as saídas. Na fase de treinamento estes algoritmos avaliam o conjunto de entrada e saída de dados conhecidos e, em seguida, nas fases de validação e teste, através de padrões aprendidos na fase anterior, recebem uma série de entrada e compara a saída prevista com a saída desejada a fim de avaliar o seu desempenho.
- **Não supervisionado:** esses algoritmos são utilizados em dados que não possuem rótulos históricos, no qual o sistema não sabe a “resposta certa”. O objetivo é descobrir o que está sendo mostrado com a exploração dos dados e encontrar alguma estrutura e ou representação;
- **Semi-supervisionado:** esses algoritmos são utilizados para as mesmas aplicações que o aprendizado supervisionado, porém eles utilizam tanto dados rotulados quanto não rotulados para o treinamento; e
- **Por reforço:** esses algoritmos descobrem pela tentativa e erro quais ações geram as maiores recompensas, sendo constituído de três componentes principais: o agente (o aluno ou tomador de decisões), o ambiente (tudo com o qual o agente interage) e as ações (o que o agente pode fazer).

Devido ao grande volume de dados não estruturados disponíveis em vários formatos, aplicações, plataforma e principalmente na Internet, surge uma necessidade de aprendizagem em grandes quantidades de dados de uma forma não-supervisionada (dados não anotados), com a realização de abstrações complexas dos dados através de um processo de aprendizagem hierárquica muito similar ao que ocorre no cérebro humano, conhecida como aprendizagem profunda (*deep learning*).

Os algoritmos de *deep learning* conseguem obter uma representação complexa e abstrata dos dados de forma hierárquica usando várias camadas de processamento não-linear. Eles apresentam uma abordagem inovadora, no sentido em que dispensam grande parte deste pré-processamento ao gerar de forma automática propriedades invariantes nas suas camadas de representação hierárquicas.

Serão avaliados nesta pesquisa os algoritmos tradici-

onais de aprendizado supervisionado Árvore de Decisão (AD J48), Naive Bayes (NB), Máquina de Suporte de Vetores (SVM) e *K-Nearest Neighbor* (KNN) e, o algoritmo de aprendizado não supervisionado de *deep learning* Máquina Restrita de Boltzmann (RBM).

### 3.2.1 Árvore de Decisão (AD J48)

A AD J48 (Sayad, 2016) é um algoritmo que utiliza o método divisão e conquista para aumentar a capacidade de predição das árvores de decisão. Assim, sempre utiliza o melhor passo avaliado localmente, sem se preocupar se esse passo vai produzir a melhor solução. Desta forma, um problema é dividido em vários sub-problemas sendo criadas sub-árvores entre a raiz e as folhas. O resultado final é uma árvore com nós de decisão e nós de folha. O nó superior de decisão corresponde ao melhor preditor denominado de nó raiz que pode manipular dados tanto categóricos quanto numéricos.

### 3.2.2 Naive Bayes (NB)

O NB (Pearl, 1988) é um algoritmo de representação do conhecimento que trabalha com o conhecimento incerto e incompleto por meio do Bayes com os pressupostos de independência entre os preditores. Uma NB, também chamada de rede probabilística ou rede causal, pode ser vista como um modelo que utiliza teoria dos grafos, condições de Markov e distribuição de probabilidades para representar uma situação, suas variáveis e estados e então realizar inferências. A condição de Markov afirma que as variáveis não-descendentes não fornecem informações adicionais sobre a variável em questão.

### 3.2.3 Máquinas de Suporte de Vetores (SVM)

As SVM foram desenvolvidas por Vapnik (Vapnik, 2000) e têm a capacidade de resolver problemas de classificação e regressão, adquirindo com o aprendizado, na etapa de treinamento, a capacidade de generalização. Considerando um problema de classes binárias, o objetivo é separar as instâncias destas duas classes através de uma função que será obtida a partir dos exemplos conhecidos na fase de treinamento. Busca-se produzir um classificador que funcione de forma adequada com exemplos não conhecidos, ou seja, exemplos que não foram aplicados durante o treinamento, adquirindo assim a capacidade de prever as saídas de futuras novas entradas.

### 3.2.4 K-Nearest Neighbor (KNN)

O KNN (Zizka et al., 2019) é um algoritmo formado por vetores  $n$ -dimensionais e cada elemento deste conjunto representa um ponto no espaço  $n$ -dimensional, onde para determinar a classe de um elemento que não pertença ao conjunto de treinamento. A finalidade é procurar  $K$  elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância, verificando quais são as classes desses  $K$  vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido. A distância pode, em geral, ser qualquer medida

métrica: a distância euclidiana padrão é a escolha mais comum. Os métodos baseados em vizinhos são conhecidos como métodos de aprendizado de máquina não generalizantes, uma vez que eles simplesmente “lembram” todos os seus dados de treinamento transformados em uma estrutura de indexação rápida.

### 3.2.5 Máquina Restrita de Boltzmann (RBM)

As RBM (Hinton and Salakhutdinov, 2006) são redes neurais de aprendizado não-supervisionado. Estas são caracterizadas principalmente por sua capacidade de aprenderem representações internas e de resolverem problemas combinatórios complexos. A camada visível corresponde aos componentes de um exemplo de entrada como, por exemplo, os *pixels* de uma imagem. A camada escondida modela a dependência entre os componentes da camada de entrada, que por sua vez deverá aprender a extrair características desses dados.

As conexões entre neurônios são bidirecionais e simétricas. Isso significa que existe tráfego de informação em ambos os sentidos da rede. Nessa topologia só existem conexões entre neurônios de camadas diferentes, justificando portanto a denominação máquina restrita.

Em sua topologia existem quatro hiper-parâmetros: quantidade de neurônios da camada visível ( $v$ ), quantidade de neurônios da camada escondida ( $h$ ), a taxa de aprendizado ( $lr$ ) e a quantidade de ciclos ( $ep$ ).

Uma taxa de aprendizado muito baixa torna o aprendizado da rede muito lento, ao passo que uma taxa de aprendizado muito alta provoca oscilações no treinamento e impede a convergência do processo de aprendizado. Geralmente seu valor varia de 0,1 a 1,0.

O número de ciclos é o número de vezes em que o conjunto de treinamento é apresentado à rede. Um número excessivo de ciclos pode levar a rede à perda do poder de generalização (*overfitting*). Por outro lado, com um pequeno número de ciclos, a rede pode não ser capaz de modelar o comportamento geral do sistema (*underfitting*).

## 3.3 Métricas de Avaliação de Desempenho

As métricas de desempenho tem objetivo de realizar a avaliação da qualidade dos classificadores (preditores) baseados em modelos de aprendizado de máquina, sendo constituídas de fórmulas matemáticas e estatísticas (Zhang and Zhou, 2014). As métricas mais utilizadas são a Acurácia, Revocação, Precisão,  $F_1$ -score e Especificidade assim detalhadas:

- **Acurácia:** é a quantidade de amostras positivas ( $AP$ ) e negativas ( $AN$ ) classificadas corretamente dividido pelo total de amostras ( $TA$ ) da série avaliada, representada pela Eq. (1).

$$\text{Acurácia} = \frac{AP + AN}{TA} \quad (1)$$

- **Precisão:** é a quantidade de amostras positivas ( $AP$ ) classificadas corretamente sobre o total de amostras classificadas como falsas positivas ( $FP$ ) acrescida das

amostras positivas em percentual, representada pela Eq. (2).

$$\text{Precisão} = \frac{AP}{FP + AP} \quad (2)$$

- **Recall:** é a quantidade de amostras positivas (AP) classificadas corretamente sobre o total de amostras classificadas como falsas negativas (FN) mais AP em percentual, representada pela Eq. (3).

$$\text{Recall} = \frac{AP}{FN + AP} \quad (3)$$

- **F1-score:** é a média harmônica entre a Precisão e o Recall, representada pela Eq. (4).

$$F1\text{-score} = \frac{2 * \text{Precisão} * \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (4)$$

- **Especificidade:** é a quantidade de amostras negativas identificadas corretamente (AN) sobre o total de amostras negativas (TAN), representada pela Eq. (5).

$$\text{Especificidade} = \frac{AN}{TAN} \quad (5)$$

### 3.4 Métricas de Avaliação de Risco

As métricas de risco (Ernst, 2014) tem a finalidade de avaliar o risco versus retorno das estratégias de negociações proposta utilizando. Serão utilizados os seguintes indicadores: volatilidade, índice Sharpe e índice Sortino.

- **Volatilidade:** é a medida da taxa de variação de um ativo num determinado período de tempo, ou seja, é quanto este ativo variou em (%) num determinado período de tempo. Onde  $P_i$  é o preço de fechamento no período  $i$  e  $P_{i-1}$  é o mesmo preço de fechamento no período anterior e  $\ln$  é o logaritmo normal, conforme Eqs. (6) e (7). O valor 252 é o período anualizado em dias úteis.

$$r = \ln\left(\frac{P_i}{P_{i-1}}\right) \quad (6)$$

$$\text{Volatilidade} = \left(\sqrt{\frac{\sum |r - \bar{r}|^2}{n}}\right) * \sqrt{252} \quad (7)$$

- **Índice Sharpe:** é uma medida de eficiência da relação do risco X retorno, ou seja, quanto maior é o valor do índice Sharpe, mais eficiente é o retorno financeiro (RF) encontrado na estratégia de operação. Para o cálculo deste índice é subtraído do RF do percentual de ganho do Ibovespa, que é um índice de renda variável, do período avaliado e o resultado dividido pela volatilidade conforme a Eq. (8).

$$\text{Sharpe} = (RF - \text{Ibovespa})/\text{volatilidade} \quad (8)$$

- **Índice Sortino:** é uma adaptação do índice de Sharpe, na qual a volatilidade do ativo é substituída pela *volatilidadeDownside*, ou seja, utilizando apenas os retornos abaixo de um certo valor de referência nas Eqs. (6) e (7). A vantagem do índice de Sortino com relação ao de Sharpe é que ele reflete apenas a volatilidade "ruim", ou seja, das perdas, conforme Eq. (9)

$$\text{Sortino} = (RF - \text{Ibovespa})/\text{volatilidadeDownside} \quad (9)$$

## 4 Metodologia

Esta Seção descreve as 5 etapas (Coleta de Dados, Transformação de Dados, Modelos de Previsão de Tendências, Estratégias de Operação e Análise de Resultados) do arcabouço metodológico implementado responsável pela geração das estratégias de operação propostas para aplicações no mercado financeiro, conforme Fig. 1.

### 4.1 Coleta de Dados

Para desenvolver a presente pesquisa utilizou-se ativos listados na Bolsa de Valores de São Paulo (B3)<sup>1</sup>, coletados por meio do terminal *Bloomberg*<sup>2</sup> que disponibiliza informações, dados, indicadores e notícias do mercado financeiro e de capitais. O acesso a esta plataforma foi disponibilizado pelo Laboratório do Grupo de Finanças (GFin) do Centro Federal de Educação Tecnológica de Minas Gerais Campus VI<sup>3</sup>.

Os dados coletados são séries históricas dos preços ajustados de preços (fechamento, abertura, máximo, mínimo, fechamento anterior), além da quantidade de negócios, quantidade de papéis, volume financeiro, data e hora das transações realizadas dos códigos e ativos avaliados.

### 4.2 Transformação de Dados

A etapa de transformação de dados é responsável por realizar a criação do atributo indicador alvo (*ClasseY*) e de normalizar cada atributo de preços (fechamento, abertura, máximo, mínimo, fechamento anterior), quantidade de negócios, quantidade de papéis e volume financeiro coletado das cotações históricas dos códigos de ativos da B3.

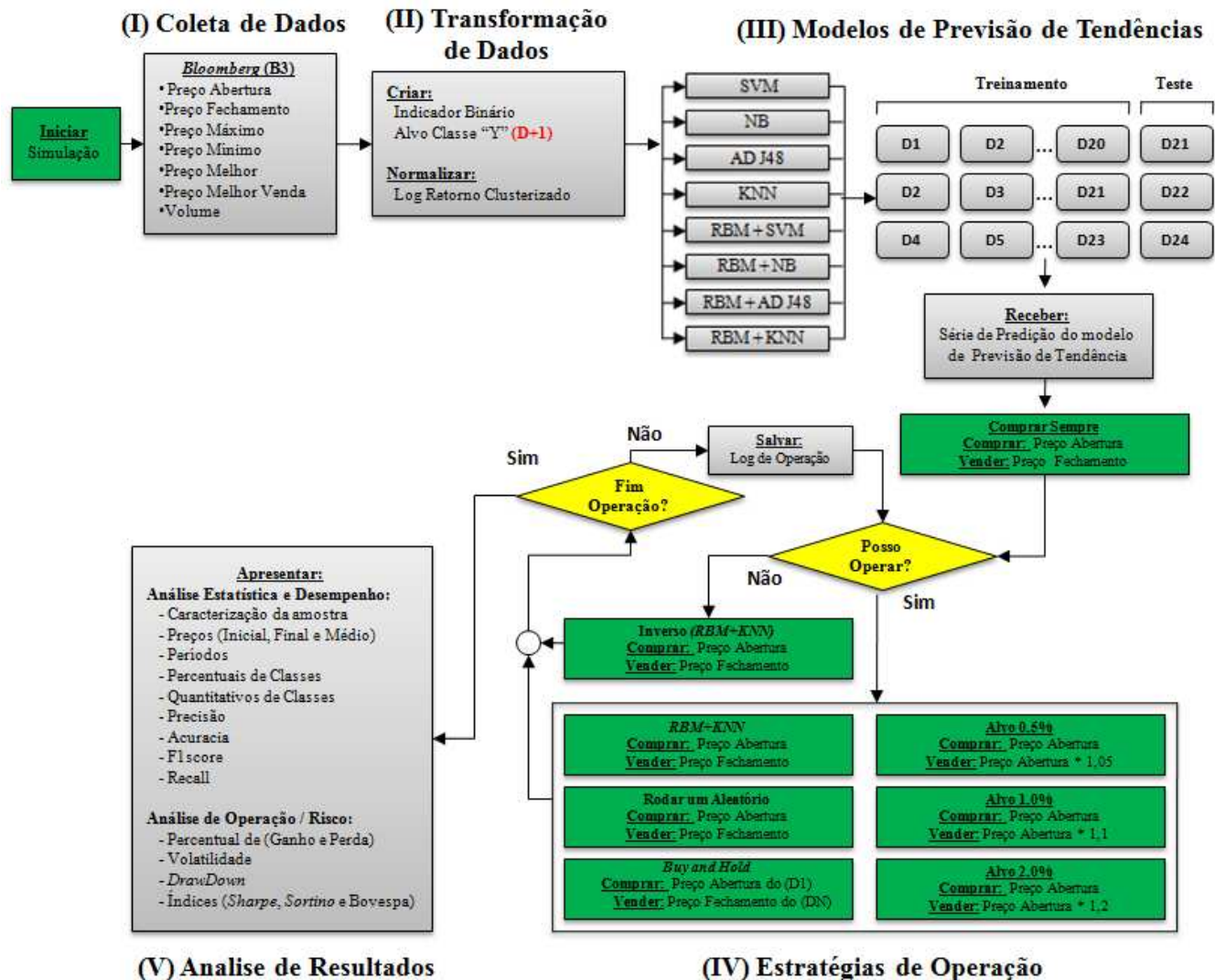
O atributo indicador alvo *ClasseY* é calculado através da diferença entre os preços de Fechamento e Abertura do próximo dia ( $D + 1$ ) da série temporal avaliada. Caso este resultado seja maior do que 0, o atributo alvo *ClasseY* recebe o valor de 1. Caso contrário, este atributo alvo recebe o valor de 0.

<sup>1</sup>Bolsa Brasil Balcão (B3): [www.b3.com.br](http://www.b3.com.br)

<sup>2</sup>Bloomberg: [www.bloomberg.com.br](http://www.bloomberg.com.br)

<sup>3</sup>GFIN: [www.dcsa.cefetmg.br/pesquisa/grupos-cnpq/gfin](http://www.dcsa.cefetmg.br/pesquisa/grupos-cnpq/gfin)





**Figura 1:** Metodologia implementada em um arcabouço de algoritmos de aprendizado de máquina para avaliação de estratégias de investimento para o mercado de ações.

Segundo Christie-David (2002), a maioria dos estudos financeiros, ao invés de utilizarem os preços dos ativos, optam por utilizar os retornos gerados por eles, devido as séries de retorno possuírem propriedades estatísticas mais atraentes, com menor variação aleatória e minimização de ruídos, facilitando desta forma, a tarefa de aprendizado e previsão dos classificadores.

Assim sendo, realiza-se a transformação das séries temporais em uma série de retorno de  $\log$  natural ( $X$ ), através da diferença entre preço inicial ( $P_i$ ) e o preço anterior ( $P_{i-1}$ ), conforme a Eq. (10).

$$X = \log(P_i) - \log(P_{i-1}) \quad (10)$$

Para retirar os ruídos (*outliers*) da série de retorno de  $\log$  natural ( $X$ ) e, conseqüentemente facilitar o trabalho dos modelos de previsão, agrupamos (*clustering*)

esta mesma série temporal ( $X$ ) em  $\alpha$  partes para que estas várias amostras de ( $\mu$ ) e ( $\sigma$ ) possam representar a série em questão avaliada, sendo detalhado o cálculo na Eq. (11).

$$\frac{x - (\sum_{j=i}^{\alpha} \mu)}{\sum_{j=i}^{\alpha} \sigma} \quad (11)$$

Quando uma série temporal de preço ( $X$ ) é muito grande, a sua média ( $\mu$ ) e o seu desvio padrão ( $\sigma$ ) podem não representar a realidade. Desta forma, foi realizada simulações variando o  $\alpha$  entre [5;15], no qual o valor 7 foi o mais adequado às séries temporais avaliadas.

### 4.3 Modelo de Previsão de Tendências

Para a etapa de classificação, inicialmente utilizou-se 4 algoritmos de aprendizado de máquina tradicionais da literatura (*SVM*, *NB*, *KNN* e *AD J48*) atuando como classificadores da categoria supervisionada para a previsão de tendência das séries avaliadas.

Em seguida, foram propostos mais 4 modelos *RBM + SVM*, *RBM + NB*, *RBM + KNN* e *RBM + AD J48* combinando inicialmente o algoritmo *RBM* de *deep learning* de categoria não-supervisionada com a finalidade de realizar redução de dimensionalidade, através de seleção e extração de características latentes e ocultas nos dados iniciais, produzindo uma saída reduzida que será utilizada como entrada para os 4 algoritmos de categoria supervisionada, totalizando desta forma, 8 propostas de modelos de previsão de tendências.

Durante a avaliação dos modelos de previsão de tendências, serão utilizados dois tipos de janelas (fixa e deslizante) de classificação para as fases de aprendizado (treinamento) e comprovação (teste) nas séries temporais. Estas janelas serão configuradas na [Seção 5](#) para verificar qual é a mais adequada em relação aos dados avaliados.

Para cada um dos 8 modelos de previsão de tendências propostos, será gerado uma nova série temporal contendo os sinais (0 = *caiu* e 1 = *subiu*) de saída da previsão de tendências dos períodos avaliados, que utilizaremos como indicadores de tendência para a próxima etapa de estratégias de operação.

### 4.4 Estratégias de Operação

Ao identificar uma oportunidade de negociação, todas as estratégias propostas realizarão ordens de compra e venda (gatilhos) diários, para serem iniciados e finalizados no mesmo dia. Após receber a série de previsão do modelo de previsão de tendências, [Fig. 1](#), são executadas 8 estratégias de operação, que serão detalhadas a seguir:

- 1) **RBM + KNN**: caso o sinal da previsão da série temporal avaliada gerada pelo modelo de previsão seja igual a 1, realiza-se a ordem de compra com o valor do Preço de Abertura do dia seguinte (D+1) e, a ordem de venda com o valor do Preço de Fechamento do dia seguinte (D+1). Caso o sinal de previsão seja igual a 0, o gatilho será cancelado;
- 2) **Comprar Sempre**: indiferente do sinal da série de previsão, será realizado uma ordem de compra com o valor do Preço de Abertura e uma ordem de venda com o valor do Preço de Fechamento todos os dias da série avaliada;
- 2) **Buy and Hold**: será realizado a uma ordem de compra com o valor do Preço de Abertura do 1º dia e uma ordem de venda com o valor do Preço de Fechamento do último dia da série avaliada;
- 3) **Aleatória**: caso o resultado da previsão da série temporal avaliada gerada pelo modelo de previsão seja igual a 1, executado um algoritmo que gera uma sequência aleatória binária (0,1) e caso o resultado seja igual a 1, realiza-se a ordem de compra com o valor

do Preço de Abertura do dia seguinte (D+1) e, a ordem de venda com o valor do Preço de Fechamento do dia seguinte (D+1). Caso contrário, o gatilho será cancelado;

- 4) **Alvos (0,5%, 1,0% e 2,0%)**: Caso o sinal da previsão da série temporal avaliada gerada pelo modelo de previsão seja igual a 1, realiza-se a ordem de compra com o valor do Preço de Abertura do dia seguinte (D+1) e, a ordem de venda com o valor do Preço de Abertura do dia seguinte (D+1) reajustado com cada um dos 3 (três) tipos de alvos simultâneos. Caso o sinal de previsão seja igual a "0", o gatilho será cancelado; e
- 5) **Inversa**: caso o sinal da previsão da série temporal avaliada gerada pelo modelo de previsão seja igual a 0, realiza-se a ordem de compra com o valor do Preço de Abertura do dia seguinte (D+1) e, a ordem de venda com o valor do Preço de Fechamento do dia seguinte (D+1). Caso o sinal de previsão seja igual a "1", o gatilho será cancelado.

Nesta etapa, serão armazenados todos os dados e indicadores que compõem o processo de execução das ordens de compra e vendas, como por exemplo, a quantidade de gatilhos e o percentual (%) (ganho, perda e retorno) acumulados.

### 4.5 Análises de Resultados

A etapa da análise de resultados apresenta dois tipos de análises distintas consolidadas para cada código de ativo avaliado: as análises (estatística /performance e de operação/risco).

A análise estatística e *performance* apresenta a caracterização da amostra contendo a distribuição dos dados em relação aos períodos de treino e teste, quantidades de dias avaliados, preços (inicial, médio e final), quantidade e percentual de dias que apresentaram altas e baixas. Além disso, realiza a avaliação dos modelos de aprendizado de máquina através das métricas previamente descritas (acurácia [Eq. \(1\)](#), *recall* [Eq. \(3\)](#), precisão [Eq. \(2\)](#) e especificidade [Eq. \(5\)](#)) com a finalidade de medir as taxas de desempenho dos classificadores. Já a análise de operação/risco é responsável por apresentar os resultados acumulados em percentual (%) e em reais (R\$) em relação ao custo operacional, retornos bruto e líquido, índices *Sharpe* e *Sortino*.

Desta forma, este arcabouço metodológico proposto avalia a viabilidade, o grau de risco de investimento das propostas de estratégias de operação utilizando técnicas de *machine learning* e, a possibilidade de aplicação futura destas estratégias em um cenário real no mercado financeiro.

## 5 Simulação e Resultados

Será realizada a calibração dos parâmetros dos algoritmos utilizados nos modelos de previsão de tendências propostos, avaliando o tipo de janela de classificação mais adequado à amostra de dados, o modelo de previsão que apresenta os melhores percentuais de



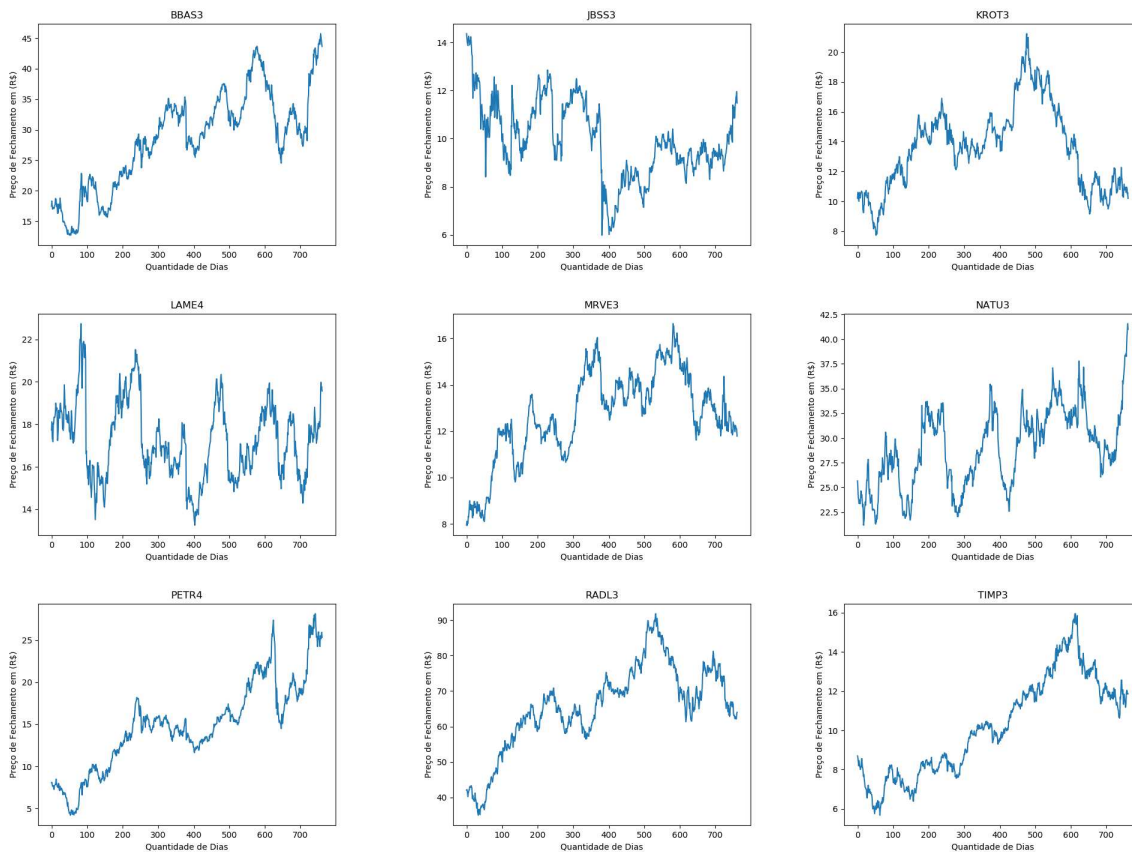


Figura 2: Série de Preço de Fechamento de 741 dias úteis no período de 03/12/2015 a 03/12/2018.

Tabela 2: Valor de Mercado e Variação dos Códigos de Ativos entre Dezembro de 2015 a Dezembro de 2018.

Ativo	Nome	Sector	Preço (DEZ 2015)	Preço (DEZ 2018)	Variação
RADL3	Raia Drogasil	Drogaria	R\$ 40,06	R\$ 63,46	58,41%
KROT3	Kroton Educacional	Educação	R\$ 10,06	R\$ 10,59	5,27%
MRVE3	MRV Engenharia	Construção Civil	R\$ 8,60	R\$ 12,06	40,23%
BBAS3	Banco do Brasil	Financeiro	R\$ 17,40	R\$ 44,10	153,45%
PETR4	Petróleo Brasileiro SA	Óleo e Gás	R\$ 7,99	R\$ 25,94	224,66%
NATU3	Natura	Cosmético	R\$ 22,50	R\$ 41,57	84,76%
JBSS3	JBS S.A.	Alimentício	R\$ 12,58	R\$ 11,96	-4,93%
LAME4	Lojas Americanas	Varejo	R\$ 18,43	R\$ 19,77	7,27%
TIMP3	TIM Brasil e Participações	Telefonia	R\$ 7,55	R\$ 11,86	57,09%

retorno financeiro e os custos operacionais existentes.

Os códigos de ativos que foram contemplados nos experimentos são de empresas que compõem o Índice Bovespa (Ibovespa), por possuírem uma maior representatividade em relação ao volume negociado, representando diferentes setores da economia como petróleo e gás, mineração, bancos, siderurgia, alimentação, educação, varejo, engenharia e construção, cosméticos, farmacêutico e telefonia, conforme detalhado na Tabela 2.

## 5.1 Avaliação da calibração dos algoritmos

Os algoritmos *Naive Bayes* (NB), Máquina de Vetor de Suporte (SVM), Árvore de Decisão (AD J48), *K-Nearest Neighbour* (KNN) e Máquina Restrita de Boltzmann (RBM) de aprendizado de máquina que serão avaliados possuem alguns parâmetros a serem calibrados que podem influenciar diretamente no resultado da classificação, que serão detalhados a seguir:

- NB: não possui parâmetros para calibração;
- SVM: testou-se duas diferentes funções *kernel* ( $k$ ) a *Radial Basis Function* (RBF) e a *polynomial* (Poly) com uma quantidade de interação ( $q_i$ ). Na função RBF devemos calibrar a largura da curva gaussiana ( $s$ ) e

o custo ( $c$ ), que define a suavização das margens do hiperplano divisor. Já na função *Poly* basta ajustar a ordem do polinômio ( $op$ ) usado;

- *ADJ48*: o  $N$  que indica o número mínimo de elementos permitidos em cada folha e o  $C$ , que é o *confidence* e escolhe o fator de confiança inicial para podar a árvore;
- *KNN*: foram avaliados diversos valores para o número de vizinhos próximos ( $k$ ) e tamanho de folha (*TF*) com ponderação de pesos uniforme e distância euclidiana entre os pontos; e
- *RBM*: testaremos diversos valores variando o quantidade de componentes ( $c$ ), a quantidade de épocas ( $q$ ) e a taxa de aprendizado ( $lr$ ) com 10 camadas ocultas e a inicialização da rede de forma determinística.

Neste sentido, após várias simulações com alternância de diversos parâmetros com o objetivo de ajustar os algoritmos aos dados em questão, os mesmos foram calibrados de acordo com a [Tabela 3](#), na qual a coluna (Resultado da Calibragem) apresenta a configuração final utilizada nos próximos experimentos.

## 5.2 Avaliação da janela de classificação

Foram utilizados dois tipos de janelas de classificação durante as fases de aprendizado (treinamento) e comprovação (teste) nas séries temporais.

Após a etapa da calibração dos algoritmos de aprendizagem de máquina, realizou-se várias simulações contemplando os 9 códigos de ativos de setores diversos, para os 3 períodos de avaliação (30, 132 e 243) de dias úteis, com o intervalo de 5 anos (2012 a 2016).

Durante as simulações variou-se a janela deslizante entre 10 e 45 dias, obtendo a configuração de 20 dias como melhor parâmetro encontrado, utilizando como peso, o percentual de precisão do modelo de previsão de tendências acima de 50%.

Já para a janela fixa, utilizou-se 3 variações (70% / 30%, 80% / 20% e 85% / 15%) para os dados (Treinamento/Teste) respectivamente, obtendo a configuração de 70% / 30% como melhor parâmetro encontrado, utilizando como peso, o percentual de precisão do modelo de previsão de tendências acima de 50%.

No entanto, ao comparar o melhor resultado de configuração entre ambas janelas (deslizante e fixa), a janela de classificação que apresentou os melhores percentuais de precisão acima de 50% para os 8 modelos de previsão de tendências proposto foi a janela deslizante contendo 20 dias (Treinamento) e 1 dia (Teste), contrariando os vários exemplos encontrados nos trabalhos relacionados ([Liu, 2017](#), [Lee, 2017](#), [Barboza et al., 2017](#), [Dixon et al., 2016](#), [Xiong, 2016](#), [Jia, 2016](#), [Hrasko, 2015](#), [Liu, 2015](#), [Ding et al., 2015](#), [Cao et al., 2015](#), [Sharang, 2015](#)).

Devido às séries temporais financeiras passarem por constantes mudanças e serem muito influenciadas por fatores externos de toda espécie (políticos, locais, macroeconômicos, dentre outros), dificilmente modelos de previsão de tendência utilizando janelas de treinamento muito grandes conseguirão generalizar e obter resultados que efetivamente representem a precisão de

seu comportamento.

Desta forma, com a finalidade de filtrar e diminuir o espaço de exploração de possibilidades de combinações com os modelos de previsões de tendência propostos, descartamos a utilização da janela de classificação fixa com 70% (Treinamento) e 30% de (Teste) para as demais simulações a seguir.

## 5.3 Avaliação dos modelos de previsão de tendências

Diante dos 8 modelos de previsão de tendências propostos e com o objetivo de escolher o melhor modelo que encontre a maior quantidade de retornos financeiros em percentual (%), realizou-se outra simulação contemplando novamente os 9 códigos de ativos para 3 períodos (30, 132 e 243) em dias úteis, somente com a janela deslizante de 20 dias (Treinamento) e 1 dia (Teste). A [Fig. 4](#) mostra que o modelo de previsão de tendência proposto RBM+KNN encontrou os melhores resultados, para os 3 períodos avaliados dos 9 ativos obtendo uma Precisão acima de 50%, uma acurácia média de 50% e um retorno financeiro líquido médio de 15% chegando até 25%, o que destacou este modelo de previsão dos demais.

## 5.4 Avaliação da viabilidade dos custos operacionais

Para todas as simulações realizadas foram contabilizados os custos das operações, contemplando o Imposto de Renda (IR) recolhido na fonte pela Bovespa, o Imposto de Renda recolhido pelo investidor sobre o lucro, a Taxa de Corretagem sobre as ordens de compra e venda, o Imposto Sobre Serviços (ISS) sobre o valor da taxa de corretagem e os Emolumentos sobre o valor investido. Devido muitas corretoras isentarem seus clientes da taxa de custódia, esta não foi acrescentada nos custos das transações.

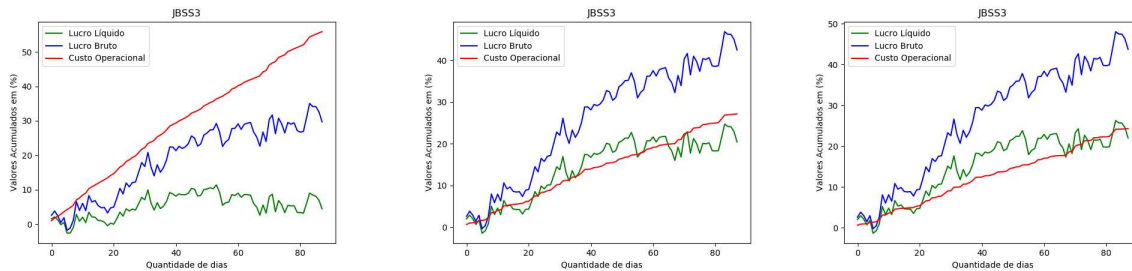
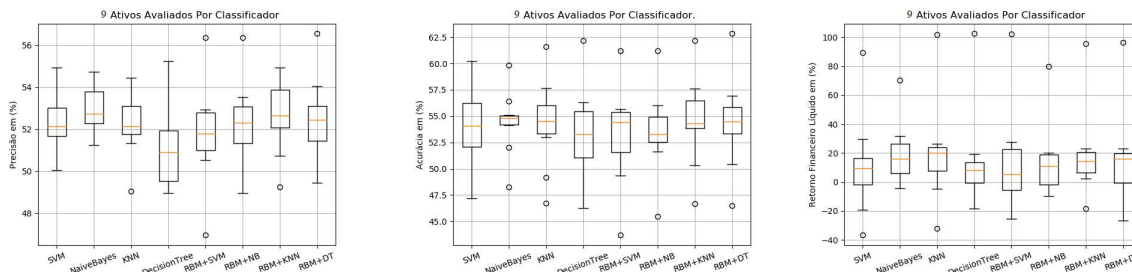
A [Fig. 3](#) apresenta um comparativo entre a mesma série temporal de 132 dias úteis no período de 01/12/2014 a 16/07/2015 do código de ativo JBSS3. Porém, contendo lotes de volume de negociação de 100, 1.000 e 10.000 respectivamente, por ordem executada. Fica evidente que o investidor com um capital menor possui um custo operacional maior o que, muitas vezes, pode inviabilizar uma estratégia de negociação lucrativa. Diante deste fator, escolhemos para as simulações um volume negociado de 1.000 lotes por estratégia de operação que for executada.

## 5.5 Avaliação da série de 741 dias úteis.

As [Figs. 2, 5 e 6](#) apresentam de forma consolidada as distribuições das séries de preço de fechamento original, os percentuais acumulados de evolução de capital e os retornos em percentual das estratégias de operação respectivamente, dos 9 códigos de ativos avaliados no período de 741 dias úteis entre 03/12/2014 a 03/12/2018. Para complementar a análise a [Tabela 4](#) apresenta de forma consolidada os resultados contendo quase 50

**Tabela 3:** Intervalo de parâmetros de calibração dos algoritmos.

Algoritmos	Parâmetros	Variação	Resultado da Calibragem
NB	-	-	-
SVM	$qi, s, k$	[1;7], [100;500], [Poly, Rbf]	$qi=200, s=0.3, k=Rbf$
AD J48	$N, C$	[1; 5], [0.01; 0.25]	$N=4, C=0.20$
KNN	$k, TF$	[0;50], [0;100]	$k=5, TF=30$
RBM	$c, q, lr$	[2;10], [100;500], [0.01;0.15]	$c=4, q=200, lr=0.10$

**Figura 3:** Evolução de capital acumulados de 132 dias úteis de 01/12/2014 a 16/07/2015.**Figura 4:** Comparação entre as medidas de desempenho e o retorno financeiro entre os 8 modelos propostos para a série de 741 dias úteis de avaliação no período 03/12/2015 a 03/12/2018.

indicadores (estatísticos, desempenho, risco e retornos financeiros).

Comparando a Fig. 2 com os indicadores de distribuição de Classe 0 e 1 em (%) da Tabela 4, os códigos de ativos que apresentaram uma distribuição de classes balanceadas, ou seja, com variações de no máximo 2,00% foram RADL3, BBAS3 e TIMP3. Já os códigos de ativos KROT3, MRVE3, PETR4, NATU3 e LAME4 apresentaram uma distribuição de classes desbalanceadas, ou seja, com variações entre classes acima 2% a 11%. No entanto ao comparar a Fig. 2 como o indicador *Buy and Hold* da Tabela 4 os códigos de ativos que apresentaram tendências de alta foram RADL3, MVRE3, BBAS3, PETR4, NATU3 e TIMP3 com variações de 33,39% a 186,46% no período avaliado de 3 anos. Já os códigos de ativos que apresentaram tendências de baixa foram KROT3, JBSS3 e LAME4 com variações de -4,09 a 6,03 ficando abaixo dos principais índices de mercado de aplicações de renda fixa para o período de 3 anos.

Conforme a Tabela 4, a variação entre a perda mínima e máxima foram respectivamente: RADL3 -4,32%, KROT3 -6,52%, MRVE3-4,53%, BBAS3 -14,68%, PETR4 -11,93%, NATU3 -9,00%, JBSS3 -25,71%, LAME4 -7,58% e TIMP3 -6,49%. O ideal para

avaliar a estratégia de operação proposta é que este indicador de risco esteja com valores abaixo de 10,00%, o que ocorreu para 6 dos 9, ou seja 66,67% dos códigos de ativos.

De acordo com a Fig. 5 de evolução de capital dos 9 códigos de ativos avaliados na simulação do modelo proposto, somente 2 (PETR4 e LAME4) apresentaram retorno financeiro negativo. Para LAME4 a estratégia de operação obteve uma retorno financeiro negativo máximo de -4,0% nos primeiros 200 dias de operação e, em seguida, começou a acertar as previsões, porém ao final obteve um retorno negativo bruto de -6,18% conforme Tabela 4. Já PETR4 teve uma queda de -40% nos primeiros 50 dias de operação, em seguida obteve um ganho de quase 60% até os próximos 140 dias, onde começou novamente apresentar quedas consecutivas obtendo um retorno financeiro bruto de -37,47%. Para os demais 7 códigos de ativos a estratégia de operação gerou retornos financeiros bruto entre 21,20% BBAS3 e 103,53% NATU3. Já o retorno financeiro líquido variou entre 16,53% BBAS3 e 80,75% NATU3.

O gráfico *Boxplot* fornece informação sobre localização, dispersão, assimetria, comprimento da cauda e medidas discrepantes (*outliers*). Conforme a Fig. 6



**Tabela 4:** Resultado da série de 741 dias úteis de avaliação no período 03/12/2015 a 03/12/2018.

Retornos / Ativos	RADL3	KROT3	MRVE3	BBAS3	PETR4	NATU3	JBSS3	LAME4	TIMP3
Oráculo (%)	479,72	709,88	506,47	630,65	666,08	655,35	806,76	534,70	554,58
Retorno Bruto (%)	64,49	74,02	60,92	21,20	-37,47	103,53	59,07	-6,18	42,66
Retorno Líquido (%)	50,31	57,73	47,52	16,53	-29,23	80,75	46,07	-4,82	33,28
Custo Operacional (%)	14,19	16,28	13,40	4,66	-8,24	22,78	12,99	-1,36	9,39
Buy And Hold (%)	48,48	4,37	33,39	127,36	186,46	70,35	-4,09	6,03	47,38
Comprar Sempre (%)	-20,79	-34,60	-25,97	-44,99	-92,95	-18,82	-35,70	-90,26	26,84
Inversa (%)	-85,28	-108,61	-86,89	-66,18	-55,47	-122,35	-94,77	-84,08	-15,82
Aleatória (%)	-51,28	-37,14	-26,66	-36,82	-59,56	-2,74	-66,01	7,19	-11,30
Alvo 0,5%	-2,70	15,30	23,95	20,31	-3,36	50,79	2,53	-30,61	12,35
Alvo 1,0%	22,98	35,15	59,70	-21,08	-17,67	84,37	-16,36	-4,30	50,67
Alvo 2,0%	50,67	40,65	99,12	-21,00	-43,27	95,24	-15,44	15,08	68,52
Perda Total (%)	-209,16	-279,27	-196,88	-302,31	-361,19	-240,58	-300,35	-262,34	-236,29
Ganho Total (%)	273,65	353,29	257,80	323,50	323,71	344,11	359,42	256,16	278,96
Retorno Médio (%)	0,18	0,24	0,18	0,06	-0,10	0,32	0,20	-0,02	0,12
Perda Média (%)	-1,24	-1,87	-1,24	-1,74	-1,89	-1,64	-1,85	-1,61	-1,37
Ganho Médio (%)	1,43	2,14	1,46	1,74	1,90	1,96	2,59	1,47	1,55
Qtde Perda	168	149	159	174	191	147	162	163	173
Qtde Ganho	192	165	176	186	170	176	139	174	180
Qtde Total	360	314	335	360	361	323	301	337	353
Qtde Previsão Classe 0	380	426	405	380	379	417	439	403	387
Qtde Previsão Classe 1	360	314	335	360	361	323	301	337	353
Precisão (%)	53,33	52,55	52,54	51,67	47,09	54,49	46,18	51,63	50,99
Acurácia (%)	55,00	55,27	55,68	52,43	49,73	56,35	52,30	53,92	51,22
F1 Score (%)	53,56	49,92	51,76	51,38	47,75	52,15	44,06	50,51	49,93
Recall (%)	53,78	47,55	51,01	51,10	48,43	50,00	42,12	49,43	48,91
Specificity (%)	56,14	62,09	59,75	53,72	50,90	62,11	60,49	57,99	53,49
Índice Sharpe (%)	2,39	9,73	10,57	-0,05	-9,09	7,42	7,47	-4,95	6,71
Índice Sortino (%)	3,99	16,25	17,52	-0,08	-13,01	12,77	12,29	-7,74	10,72
Média (%)	0,18	0,24	0,18	0,06	-0,10	0,32	0,20	-0,02	0,12
Desvio Padrão (%)	1,72	2,54	1,72	2,36	2,66	2,30	3,61	2,02	1,91
Desvio Absoluto (%)	1,34	2,01	1,35	1,74	1,90	1,79	2,21	1,54	1,46
Mediana (%)	0,11	0,16	0,16	0,11	-0,08	0,37	-0,18	0,11	0,08
Quantil (%)	0,11	0,16	0,16	0,11	-0,08	0,37	-0,18	0,11	0,08
DrawDown (%)	-4,32	-6,52	-4,53	-14,68	-11,93	-9,00	-25,71	-7,58	-6,49
Máximo (%)	6,16	7,91	6,13	10,00	11,06	9,93	25,73	5,67	6,05
Amplitude (%)	10,48	14,43	10,66	24,68	22,99	18,93	51,44	13,25	12,54
Variância (%)	2,95	6,44	2,96	5,58	7,06	5,31	13,06	4,06	3,65
Qtde Dias Avaliados	741,00	741,00	741,00	741,00	741,00	741,00	741,00	741,00	741,00
Preço Inicial (R\$)	40,06	10,06	8,60	17,40	7,99	22,50	12,58	18,43	7,55
Preço Final (R\$)	63,46	10,59	12,06	44,10	25,94	41,57	11,96	19,77	11,86
Distribuição Classe 0 (%)	51,69	53,17	53,44	50,88	52,63	52,50	55,47	52,50	50,34
Distribuição Classe 1 (%)	48,31	46,83	46,56	49,12	47,37	47,50	44,53	47,50	49,66
Qtde Real de Classe 0	383	394	396	377	390	389	411	389	373
Qtde Real de Classe 1	358	347	345	364	351	352	330	352	368
Volatilidade (%)	17,89	5,37	3,71	10,41	6,51	11,03	5,00	5,63	3,12

o modelo de previsão de tendências proposto (RBM + KNN) foi superior às demais estratégias de operação para todos os códigos de 9 ativos de ativos avaliados.

Devido à estratégia de operação principal ser executada somente quando a série prevista do modelo de previsão de tendências indicar que seja um dia de alta, analisaremos somente os resultados do indicador de desempenho de precisão. Somente para LAME4 e JBSS3 apresentou valores de precisão abaixo de 50%. Já para os demais 7 códigos de ativos a variação da precisão foi de 50,99% a 54,49, obtendo uma precisão média de 51,16% para os 9 ativos avaliados. No entanto, nem sempre uma taxa de precisão do modelo de previsão acima de 50% é garantia de que a estratégia de operação obtenha retorno financeiro positivo e acima dos prin-

cipais índices do mercado, viabilizando desta forma, o modelo de operação proposto.

Os índices *Sharpe* e *Sortino* apresentam valores negativos somente para os códigos de ativos BBAS3, PETR4 e LAME4, que consequentemente obtiveram o percentual de retorno financeiro líquido negativo e/ou abaixo dos principais índices de mercado que foi o caso do BBAS3, o que inviabilizou a estratégia de operação gerada pelo modelo de previsão de tendência. Para os demais 6 códigos de ativos avaliados os índices *Sharpe* e *Sortino* variaram de 2,39% a 10,57% e 3,99% a 10,57% respectivamente, confirmando o resultado positivo apresentado pelo retorno financeiro líquido destes ativos conforme a [Tabela 4](#).

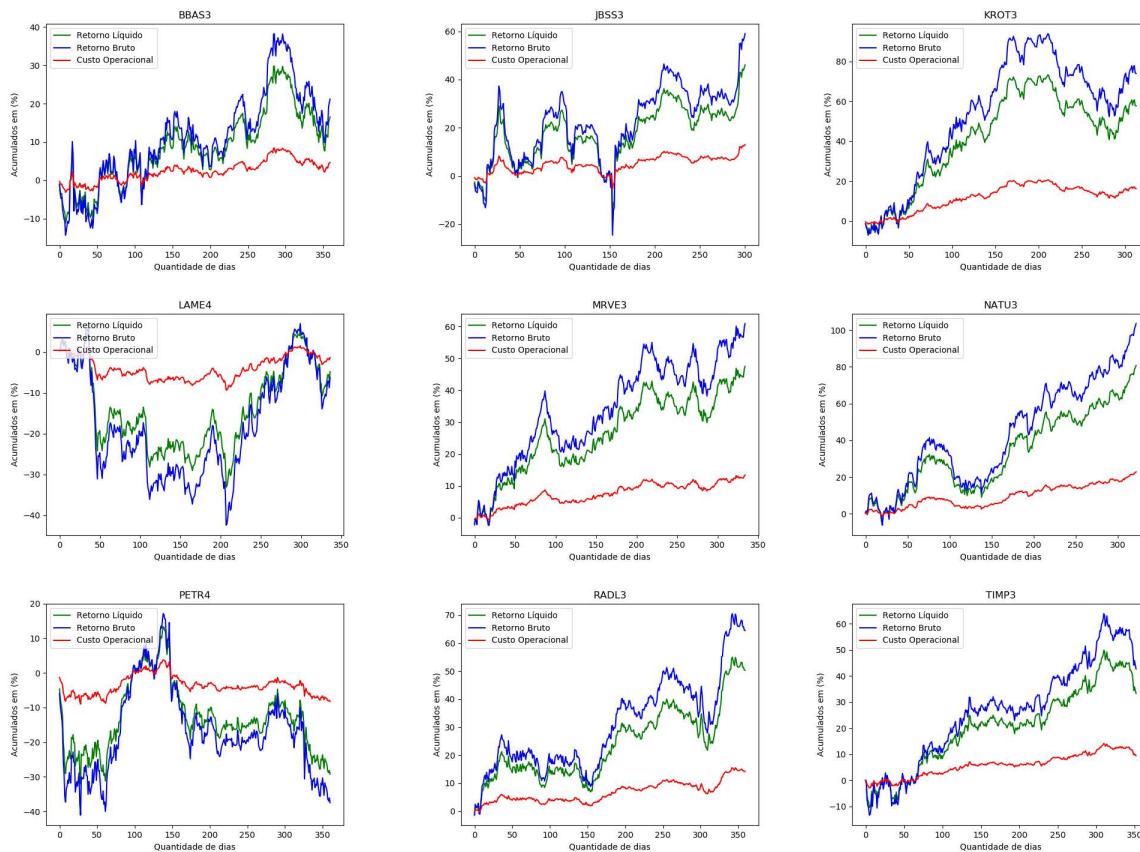


Figura 5: Evolução de capital acumulado 741 dias úteis no período de 03/12/2015 a 03/12/2018.

## 6 Conclusão e Trabalhos Futuros

Foram implementados 4 modelos de previsão (*SVM*, *NB*, *AD J48* e *KNN*) baseados em algoritmos de aprendizado de máquina. Posteriormente, foram realizadas mais 4 combinações com o algoritmo *RBM* de aprendizagem de categoria não supervisionada (*deep learning*) com os algoritmos tradicionais citados anteriormente (*RBM + SVM*, *RBM + NB*, *RBM + AD J48* e *RBM + KNN*) a fim de realizar a seleção e extração de características dos atributos envolvidos, totalizando 8 propostas de modelos de previsão de tendências.

O modelo híbrido (*RBM + KNN*) de previsão de tendências foi selecionado por ter sido considerado o mais adequado aos dados das séries temporais dos períodos avaliados para os 9 códigos de ativos. Essa adequação ocorreu porque, após vários experimentos e configurações relacionadas à calibração dos algoritmos, à escolha do tipo de janela de classificação, à verificação e seleção, o referido modelo foi o único que apresentou um percentual de retorno financeiro bruto maior do que o percentual de precisão com valores acima de 50%.

A pesquisa apresentou resultados parciais bastante interessantes com a avaliação das séries de 700 dias úteis em períodos distintos, entre os anos de 2015 a 2018. O objetivo foi de avaliar o comportamento do modelo de previsão de tendências em relação ao seu

percentual de precisão, retorno financeiro líquido e demais indicadores.

Conforme a Tabela 4, caso os 9 códigos de ativos avaliados fossem tratados como uma carteira de investimentos, teríamos um *drawdown* médio de no máximo 10,00%, um retorno financeiro bruto médio de 42,47% e líquido médio de 33,13%. Para 7 dos 9, ou seja, 77,80% dos códigos de ativos avaliados, a quantidade de acertos do modelo de previsão de tendências foram superiores à quantidade de perdas. Entretanto, somente 5 códigos de ativos, para um período de aplicação de 3 anos de avaliação, geraram retorno financeiro líquido acima de 42,00%, ou seja, acima de 14,00\$ ao ano.

Comparando os valores percentuais do oráculo em (%), que é o gabarito caso o modelo de previsão de tendência acertasse todas as previsões de alta, com o Ganho Total (%), este acertou em média 60% das previsões de alta, com Ganho Total médio entre os 9 códigos de ativos de 307,84%. Porém, as Perdas Totais (%) consumiram 86,20% do Ganho Total médio, ou seja, ainda é necessário ajustar o modelo híbrido (*RBM + KNN*) acrescentando às estratégias de operação, técnicas de *stops* e/ou medidas de risco para minimizar os prejuízos.

Verificou-se que o volume negociado, o capital disponível para a aplicação e a quantidade de gatilhos (compra e venda) influenciam diretamente no custo opera-

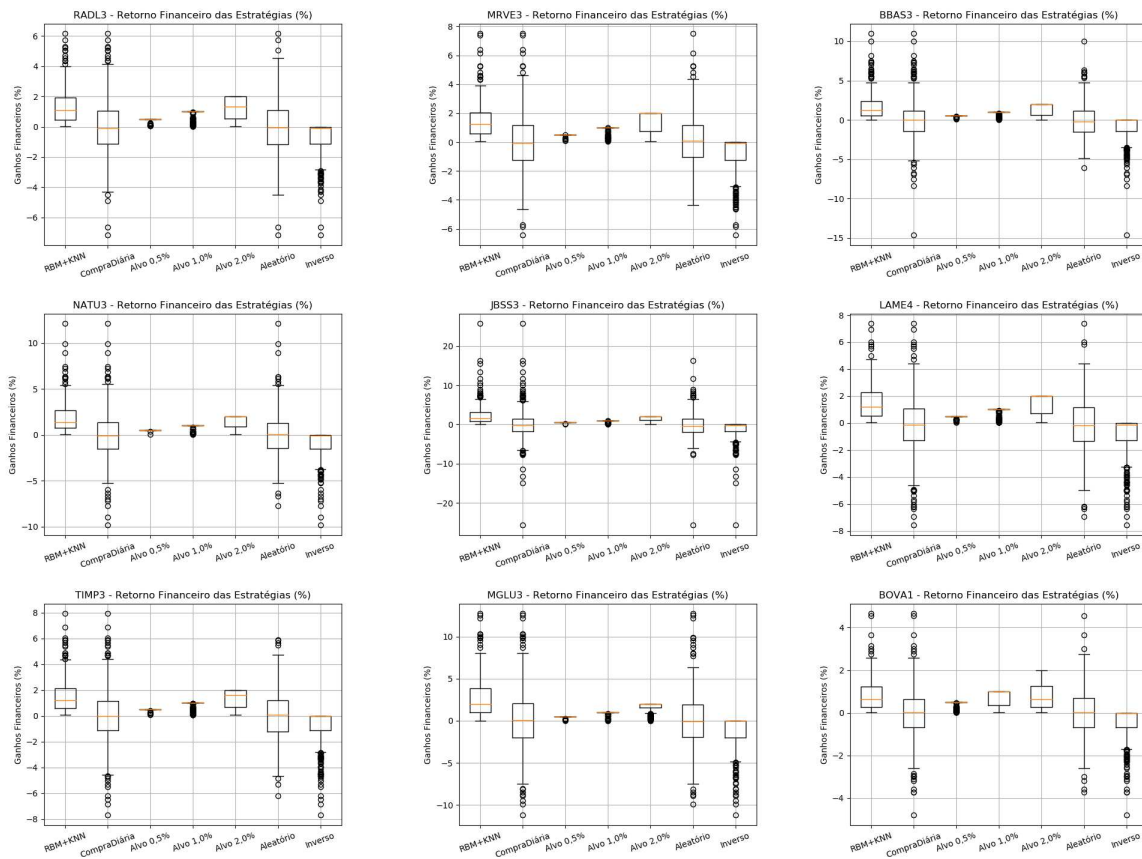


Figura 6: BoxPlot das séries de Retorno em (%) das Estratégias de 741 dias úteis no período de 03/12/2015 a 03/12/2018.

cional e podem inviabilizar a estratégia de operação, caso ela não seja configurada e adequada previamente à realidade do valor de mercado do código de ativo avaliado. É importante ressaltar que o custo operacional não é tributado de forma proporcional, pois possui algumas taxas com valores fixos. Além disso, o custo operacional é considerado muito elevado para aplicação em renda variável na bolsa de valores brasileira, dependendo do capital inicial disponível de pequenos e médios investidores, o que dificulta ainda o acesso e a popularização de investimentos neste seguimento de mercado.

Fatores externos como, notícias, os horários de funcionamento limitado entre 10:00h e 17:00h da B3, a diminuição do calendário de funcionamento da B3 por conta da grande quantidade de feriados e, principalmente, a quebra de sazonalidade na série temporal de preço com as variações percentuais significativas entre o preço de abertura do dia atual e o preço de fechamento do dia anterior, podem inviabilizar na qualidade do retorno de uma proposta de estratégia de operação implementada. Desta forma, não é possível garantir que sempre que um modelo de previsão de tendências apresentar uma precisão de desempenho acima de 50%, será encontrado também um percentual de

retorno financeiro líquido viável e acima dos *baselines* de mercado.

A principal contribuição do trabalho foi a apresentação deste modelo completo de avaliação de estratégias de negociação baseado em algoritmos de aprendizado de máquina, onde o arcabouço metodológico e suas etapas (Coletar, Transformar, Classificar, Operar e Analisar) os dados contempla todas as fases para uma análise robusta dos dados, como um recurso para a tomada de decisão de forma automatizada para investidores, que pode ajudar a reduzir riscos e maximizar o lucro.

As principais vantagens de utilização deste arcabouço metodológico proposto são:

- Geração de mais de 50 tipos de indicadores sobre os modelos de previsão de tendências suas respectivas estratégias de operação contendo informações estatísticas, desempenho, operações financeiras, dados de risco, dentre outros;
- Realização de análise temporal com a geração de forma automatizada dos gráficos de Preço de Fechamento, Evolução de capital (bruto, líquido e custo operacional), *Boxplot* do retorno financeiro líquido em (%) de cada estratégia por código de ativo;
- Possibilidade de personalizar e customizar para diferentes alvos os algoritmos e as estratégias de ope-



ração de acordo com o perfil e interesse de cada investidor; e

- Acompanhamento e simulação em paralelo da distribuição das séries temporais de diversos tipos de investimentos, como por exemplo, bolsa de valores, criptomoedas, renda fixa e outros mercados.

A disponibilização deste arcabouço contribuirá principalmente para evitar que outros estudantes e interessados da área de mercado financeiro tenham que despender muito tempo de suas pesquisas em implementar algoritmos e possam dedicar esforços em criar, validar e aprimorar suas estratégias de negociação.

Como trabalhos futuros, deseja-se comparar o modelo (RBM + KNN) proposto com as redes neurais recorrentes e *Long Short-Term Memory (LSTM)* conhecidas como o estado da arte para a previsão de séries temporais. Em seguida, utilizar outros dados de bolsas internacionais, principais índices de mercado e dados de criptomoedas. E por último, detalhar mais os indicadores da Tabela 4 caracterizando quando e como ocorrem as perdas para tentar inferir técnicas e ou estratégias de limitação de risco com o intuito de maximizar os ganhos.

## Referências

- Aamodt, T. (2015). *Predicting stock markets with neural networks*, Master's thesis, University of Oslo. Disponível em <http://urn.nb.no/URN:NBN:no-49006>.
- Assis, C., Machado, E., Pereira, A. and Carrano, E. (2018). Hybrid deep learning approach for financial time series classification, *Revista Brasileira de Computação Aplicada* 10(2): 54–63. <https://doi.org/10.5335/rbca.v10i2.7904>.
- B3-Variável, B. B. B. (2020). Tarifas de renda variável. Disponível em <http://www.b3.com.br/pt-br/produtos-e-servicos/tarifas/mercado-de-balcao/tarifas-de-renda-variavel/>.
- Barboza, F., Kimura, H. and Altman, E. (2017). Machine learning models and bankruptcy prediction, *Expert Systems with Applications* 83: 405 – 417. <https://doi.org/10.1016/j.eswa.2017.04.006>.
- Booth, A. (2016). *Automated algorithmic trading: machine learning and agent-based modelling in complex adaptive financial markets*, PhD thesis, University of Southampton. Disponível em <https://eprints.soton.ac.uk/397453/>.
- Cao, W., Hu, L. and Cao, L. (2015). Deep modeling complex couplings within financial markets, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, AAAI Press, p. 2518–2524. Disponível em <https://api.semanticscholar.org/CorpusID:8236168>.
- Chong, E., Han, C. and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies, *Expert Systems with Applications* 83: 187 – 205. <https://doi.org/10.1016/j.eswa.2017.04.030>.
- Chorafas, D. N. (1988). *Sistemas especialistas - aplicações comerciais*, McGraw-Hill Ltda.
- Christie-David, R. (2002). Analysis of financial time series, *Journal of Financial Research* 25: 445–446.
- da Silva, D. N. E. (2014). *Predição de tendência de ativos em séries temporais financeiras utilizando algoritmos de aprendizado de máquina*, Master's thesis, Instituto Militar de Engenharia.
- de Sousa, J. L. and Lopes, C. R. (2014). Aplicando técnicas de aprendizado de máquina em planejamento probabilístico, *Faculdade de Computação Universidade Federal de Uberlândia*.
- de Souza Roca, R. L. R. and Mól, A. L. R. (2015). Previsão do Índice bovespa por meio de redes neurais artificiais, *Business and Management Review*.
- Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015). Deep learning for event-driven stock prediction, *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, AAAI Press, p. 2327–2333. Disponível em <https://www.ijcai.org/Proceedings/15/Papers/329.pdf>.
- Dixon, M., Klabjan, D. and Bang, J. (2016). Classification-based financial markets prediction using deep neural networks, *Algorithmic Finance*. <http://dx.doi.org/10.2139/ssrn.2756331>.
- Ernst, A. (2014). Find the right people, processes and technology to manage record-to-report risks, *Managing Operational Tax Risk*.
- Estrada, G. B. (2015). *Deep learning for multivariate financial time series*, Master's thesis, Royal Institute of Technology.
- Ferreira, J. C. P. (2016). *Análise de influência de utilizadores e redes sociais em microblogs sobre mercados financeiros*, Master's thesis, Engenharia e Gestão de Sistemas de Informação - Escola de Engenharia - Universidade do Minho.
- Hamilton, W. P. (1922). *Stock market barometer: a study of its forecast value based on Charles H. Dow theory of the price movement. With an analysis of the market and its history since 1897*, New York: Harper & Bros.
- Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks, *Science* 313(5786): 504–507. <https://doi.org/10.1126/science.1127647>.
- Hrasko, R. (2015). Time series prediction using restricted boltzmann machines and backpropagation, *Information Technology and Quantitative Management*. <http://dx.doi.org/10.1016/j.procs.2015.07.104>.
- Hrasko, R., Pacheco, A. G. and Krohling, R. A. (2015). Time series prediction using restricted boltzmann machines and backpropagation, *Procedia Computer Science* 55: 990 – 999. <https://doi.org/10.1016/j.procs.2015.07.104>.

- Iglesias Caride, M., Bariviera, A. F. and Lanzarini, L. (2018). *Stock Returns Forecast: An Examination By Means of Artificial Neural Networks*, Springer International Publishing, Cham, pp. 399–410. [https://doi.org/10.1007/978-3-319-69989-9\\_23](https://doi.org/10.1007/978-3-319-69989-9_23).
- Jia, H. (2016). Investigation into the effectiveness of long short term memory networks for stock price prediction, *Cornell University Library*. Disponível em <https://arxiv.org/abs/1603.07893>.
- Kaupa, P. H. (2013). *Aplicação de técnica de inteligência artificial na seleção de ações para investimento na bolsa de valores de são paulo*, Master's thesis, Universidade Nove de Julho.
- Kitchenham, B. and Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Tecnical Report EBSE-2007-01.
- Lee, J. (2017). Deep learning-based corporate performance prediction model considering technical capability, *Sustainability*. <http://dx.doi.org/10.3390/su9060899>.
- Liang, Q., Rong, W., Zhang, J., Liu, J. and Xiong, Z. (2017). Restricted boltzmann machine based stock market trend prediction, *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1380–1387. <https://doi.org/10.1109/IJCNN.2017.7966014>.
- Liu, J. (2017). Restricted boltzmann machine based stock market trend prediction, *International Joint Conference on Neural Networks*. <http://dx.doi.org/978-1-5090-6182-2/17>.
- Liu, T. (2015). Deep learning for event-driven stock prediction, *International Joint Conference on Artificial Intelligence*.
- Mansouri, E. A. (2016). *Using artificial neural networks and sentiment analysis to predict upward movements in stock price*, Master's thesis, Worcester Polytechnic Institute.
- Melo, B. (2012). Considerações cognitivas nas técnicas de previsão no mercado financeiro, *Universidade Estadual de Campinas*.
- Nametala, C. A. L. (2017). *Construção de um robô investidor baseado em redes neurais artificiais e preditores econômicos*, Master's thesis, Escola de Engenharia da Universidade Federal de Minas Gerais (UFMG).
- Niaki, S. and Hoseinzade, S. (2013). Forecasting s&p 500 index using artificial neural networks and design of experiments, *Journal of Industrial Engineering International* 9(1): 1–9. <http://dx.doi.org/10.1186/2251-712X-9-1>.
- Oliveira, T. (2015). Predição do tráfego de rede de computadores usando redes neurais tradicionais e de aprendizagem profunda, *Universidade Federal de Uberlândia, UFU, Faculdade de Computação, FACOM*.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Ramalho, I. (2014). *Redes neurais artificiais na predição das principais séries do índice ibovespa e suas aplicações em sistemas automatizados de negociação*, Master's thesis, Universidade Federal do Rio de Janeiro.
- Santos, H. S. (2016). *Um estudo sobre o mercado brasileiro de ações a partir de dados do twitter*, Master's thesis, Mestrado em Ciências da Computação da Universidade Federal de Minas Gerais.
- Sayad, S. (2016). Decision tree – classification, *An Introduction to Data Mining* p. 18.
- Sharang, A. (2015). Using machine learning for medium frequency derivative portfolio trading, *Cornell University Library*. Disponível em <https://arxiv.org/abs/1512.06228>.
- Shen, F., Chao, J. and Zhao, J. (2015). Forecasting exchange rate using deep belief networks and conjugate gradient method, *Neurocomputing* 167: 243 – 253. <https://doi.org/10.1016/j.neucom.2015.04.071>.
- Vapnik, V. (2000). *The Nature of Statistic Learning Theory*, Springer, pp. 17–34. [http://dx.doi.org/10.1007/978-1-4757-3264-1\\_2](http://dx.doi.org/10.1007/978-1-4757-3264-1_2).
- Vengertsev, D. (2014). *Deep learning architecture for univariate time series forecasting*, Master's thesis, Department of Computer Science Stanford University.
- Wei, L.-Y. and Cheng, C.-H. (2012). A Hybrid Recurrent Neural Networks Model Based on Synthesis Features to Forecast the Taiwan Stock Market, *International Journal of Innovative Computing, Information and Control* 8(8): 504–507. Disponível em <https://api.semanticscholar.org/CorpusID:12343063>.
- White, H. (1988). Economic prediction using neural networks: The case of ibm daily stock returns, *Neural Networks, 1988., IEEE International Conference on*, IEEE, pp. 451–458.
- Xiong, R. (2016). Deep learning stock volatility with google domestic trends, *Cornell University Library*. Disponível em <https://arxiv.org/abs/1512.04916>.
- Zhang, M. and Zhou, Z. (2014). A review on multi-label learning algorithms, *IEEE Transactions on Knowledge and Data Engineering* 26(8): 1819–1837. <https://doi.org/10.1109/TKDE.2013.39>.
- Zhang, Y. (2015). *Using financial reports to predict stock market trends with machine learning techniques*, Master's thesis, University of Oxford.
- Zizka, J., Darena, F. and Svoboda, A. (2019). *Nearest Neighbors*, Taylor & Francis Group, pp. 163–172. <http://dx.doi.org/10.1201/9780429469275-6>.