# Proposal and Implementation of Machine Learning and Deep Learning Models for Stock Markets

Eduardo Jabbur Machado [1], Renato Oliveira [2,4], Adriano César Machado Pereira[1,2,3],

[1]Federal Center for Tech. Education of MG
Belo Horizonte - Brazil
ejabbur@gmail.com

[2]Federal University of Minas Gerais (UFMG)
Belo Horizonte - Brazil
[3]adrianoc@dcc.ufmg.br
[4]7renato@outlook.com

## ABSTRACT

*The investment market has been growing every day, performing an important role in the lives of individuals and corporations. Therefore, there is a need to better understand the situations that occur in the capital market, by means of strategies and indicators that can help in pattern recognition, analisys and investiment decisions. This work makes a study of characterization and analysis of historical time series data of 9 asset codes (i.e., BBAS3, PETR4, USIM5) of the Bovespa index with the proposal of evaluating one classification model. It proposes the combination of deep learning and machine learning computational intelligence models for prediction using KNN, RBM and LSTM, allowing the execution and cancellation of buy and sell orders. Finally, it evaluates the behavior of each proposed trading strategy by Accuracy, Percentage of Financial Return and other indicators that helps in a better understanding of financial market behavior.*

## Keywords

Web 2.0, Data Characterization, Stock Markets, Machine Learning, Deep Learning Trading Strategies, Financial Indicators.

## 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 à alguns fatores como impulso, razões, 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, técnicas e análises de mercado como a Análise Técnica que será abordada neste trabalho.

Para [11], na Análise Técnica, toda a informação sobre o comportamento do mercado financeiro é baseada em dados de preço e volume, sendo refletida e apresentada na forma de indicadores e gráficos que sinalizam o momento mais adequado para realizar estratégias de negociação no mercado de ações.


DOI: http://dx.doi.org/10.1145/2820426.2820444 .

Vários autores [6][2] [18] [13] realizaram estudos com o objetivo de apontar correlações entre séries temporais financeiras e indicadores técnicos, dados de notícias e outros dados.

Segundo [20], os agentes investidores, antes de tomarem uma decisão para a negociação de compra e venda de ações, utilizaram recursos de diversos meios de comunicação e informação, com destaque para a Web, que nos tempos atuais tornou-se um importante meio de produção de dados e divulgação de informações para o mercado financeiro.

Por exemplo, [17] aplicou redes neurais artificiais do tipo *Multilayer Perceptron* e a Teoria dos *Rough Sets* na seleção de ações para investimento na Bolsa de Valores de São Paulo, selecionando carteiras de ações para investimento e aplicando o modelo de *Markowitz* e a equação de Ponderação para o gerenciamento do risco. Por fim, esses retornos financeiros das carteiras gerenciadas foram comparados com o índice Ibovespa utilizado como *benchmark*.

Por outro lado, outros autores [4] [21] [14] [1] [16] [17] [23] dedicaram esforços no sentido de desenvolver modelos de classificação de tendências de retornos de ativos financeiros utilizando na maioria das vezes modelos 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.

Diante disso, pretende-se nesse trabalho investigar e analisar os seguintes pontos:

- Tratamento dos Dados: qual a melhor forma e ou técnicas de ciências dos dados para selecionar, preparar e analisar os dados de séries temporais de ações negociados em bolsa de valores.
- Modelo de Previsão: Como configurar, treinar, medir, combinar e selecionar modelos de classificação com o objetivo de prever tendências de retornos.
- Modelo de Operação: De que maneira podemos interpretar a série de previsões dos modelos de modo a criar estratégias de operação no mercado buscando taxas de retornos satisfatórias e que superem valores de referências - *baselines* - mais comuns.

Este trabalho propõe uma metodologia composta por 5 etapas (coleta, transformação, classificação, operação e análise de resultados), que utilizam dados históricos de cotações de ações disponibilizados pela Bovespa (preço e volume) de 3 códigos de ativos (USIM5, BBAS3 e PETR4) de segmentos distintos de mercado,

como entrada de três modelos de previsão de retornos 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 a fundamentação teórica detalhando as medidas de risco, medidas de desempenho e os algoritmos de aprendizado de máquina utilizados neste trabalho. A metodologia proposta está descrita na Seção 3, sendo composta 5 etapas (Coletar, Transformar, Classificar, Operar e Analisar) dados. A Seção 4 apresenta a simulação realizada juntamente com a apresentação dos resultados. Finalmente, a Seção 5 descreve as conclusões e direções futuras.

# 2. FUNDAMENTAÇÃO TEÓRICA

Esse trabalho utiliza fundamentalmente ferramentas de Análise Técnica que procuram identificar padrões específicos nas séries temporais bem como apontar tendências ou variações de volatilidade.

Nesta seção serão apresentadas métricas de risco, medidas de desempenho bem como os algoritmos de aprendizado de máquina utilizados como preditores nas estratégias de negociação.

## 2.1 Medidas de Risco

Com a finalidade de avaliar a relação risco/retorno [5] [22] de uma estratégia de negociação, serão utilizados indicadores de volatilidade *Downside*, índice *Sharpe* e *Sortino* detalhados a seguir:

- **Índice de Volatilidade**: é a medida da taxa de variação de um ativo num determinado período de tempo em percentual. Onde $Pi$ é o preço de fechamento no período $i$ e $Pi-1$ é o preço de fechamento no período anterior e $ln$ é o logaritmo normal, conforme Equações 1 e 2. O valor 252 é o período anualizado em dias úteis na Equação 2.

$$r = ln(\frac{Pi}{Pi-1}) \tag{1}$$

$$Volatilidade = (\sqrt{\frac{\sum_{|r-\bar{r}|}{}^2}{n}}) * \sqrt{252} \tag{2}$$

- **Índice *Sharpe***: é uma medida do risco em excesso esperado em relação a sua variabilidade ao adotar uma estratégia mais arriscada que uma outra estratégia submetido a uma chamada taxa livre de risco. Nesse trabalho foi utilizado como taxa livre de risco a taxa CDI, conforme Equação 3.

$$Sharpe = (RF - CDI)/volatilidade \tag{3}$$

- **Índice *Sortino***: é uma adaptação do índice de *Sharpe*, na qual a volatilidade do ativo é substituída pela *volatilidade-Downside*, ou seja, utilizando apenas os retornos negativos na Equação 1. A vantagem do índice de *Sortino* com relação ao de *Sharpe* é que ele reflete apenas a volatilidade "ruim", ou seja, das perdas, conforme Equação 4

$$Sharpe = (RF - CDI)/volatilidadeDownside \tag{4}$$

## 2.2 Medidas de Desempenho

As medidas de desempenho a seguir avaliam o desempenho dos modelos de previsão obtidos, sendo constituídas de fórmulas matemáticas e estatísticas [12]. As métricas mais utilizadas são a Acurácia, (*Recall*), Precisão, (*F1-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 em percentual.

$$Acurcia = \frac{AP + AN}{TA} \tag{5}$$

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

$$Recall = \frac{AP}{FN + AP} \tag{6}$$

- **Precisão**: é a quantidade de amostras positivas (AP) classificadas corretamente sobre o total de amostras classificadas como falsas positivas (FP) mais AP em percentual.

$$Precisao = \frac{AP}{FP + AP} \tag{7}$$

- ***F1-score***: é a média harmônica entre a Precisão e o *Recall* calculada pela Equação 8.

$$F1\text{-}score = \frac{2 * Precisao * Recall}{Precisao + Recall} \tag{8}$$

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

$$Especificidade = \frac{AN}{TAN} \tag{9}$$

## 2.3 Modelos de Aprendizado de Máquina

Os modelos de inteligência artificial são divididos em dois grupos específicos: *machine learning* e *deep learning*. Os modelos de (*Machine Learning*) são subdivididos em quatro grupos de algoritmos de aprendizado: Supervisionado, Não Supervisionado, Semi-Supervisionado e Por Reforço. Já os modelos de *deep learning* são redes neurais específicas profundas e complexas capazes de lidar com a classificação de conjuntos de dados gigantes, como por exemplo, reconhecimento de voz, imagens e obstáculos de carros autônomos [3].

Os modelos de *machine learning* de aprendizado Supervisionado são algoritmos que na fase de treinamento avaliam o conjunto de entrada e saída de dados conhecidos. A partir disso, o referido modelo compara a saída prevista com a saída desejada, a fim de avaliar o desempenho calibrando os pesos sinápticos do modelo de predição, utilizando padrões para prever os valores do rótulo em dados adicionais não rotulados nas fases de validação e teste[15].

Já os modelos de *deep learning* aplicam os conceitos do aprendizado Não Supervisionado, onde possuem disponíveis somente os padrões de entrada para a rede e, desenvolve-se nela uma habilidade de formar representações internas para codificar características de entrada e criar novas classes ou grupos automaticamente. Este tipo de aprendizado só é possível quando existe redundância e um grande volume de dados de entrada [19].

Neste trabalho, serão utilizados o algoritmo de *machine learning* de categoria Supervisionada (*K-Nearest Neighbor (KNN)*) e, os al-

goritmos de *deep learning* Máquina Restrita de *Boltzmann (RBM)* e *Long-Short Term Memory (LSTM)*).

### 2.3.1 *K-Nearest Neighbor (KNN)*

O KNN é um modelo de aprendizado utilizado geralmente como um classificador, 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 [10].

O *KNN* procura $K$ elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância. Em seguida, verifica-se quais são as classes desses K vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido.

O princípio por trás dos métodos do *KNN* é encontrar um número pré-definido de amostras de treinamento mais próximas da distância do novo ponto e prever o rótulo a partir delas. O número de amostras pode ser uma constante definida pelo usuário $K$ ou variar com base na densidade local de pontos (aprendizado de vizinho baseado em raio).

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.

### 2.3.2 *Máquina Restrita de Boltzmann (RBM)*

As *RBM* [8] 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 [8].

Na *RBM* as conexões entre neurônios são bidirecionais e simétricas. Isso significa que existe trafego de informação em ambos os sentidos da rede. Nessa topologia só existem conexões entre neurônios de camadas diferentes, justificando portanto a denominação máquina restrita [8].

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 [7]. Já 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*) [7].

### 2.3.3 *Long-Short Term Memory (LSTM)*

As redes neurais *LSTM* [9] são um tipo de rede neural recorrente (*Recurrent Neural Network (RNN)*), ou seja, uma rede capaz de processar dados sequenciais no tempo. As *RNN* implementam mecanismos de memória por meio de laços de retroalimentação entre a saída da rede e a sua entrada. A presença desses laços de retroalimentação é que possibilita esse tipo de rede neural utilizar a dimensão do tempo para associar a uma determinada entrada no tempo $t$ uma saída correspondente no tempo $k$ posterior a $t$. As redes *LSTM* são bastante utilizadas atualmente em aplicações de tradução, reconhecimento de fala e escrita e análise de sentimento o que também tem suscitado a pesquisa para sua utilização no processamento de séries temporais financeiras.

As *RNN* comuns utilizam o método de treinamento de retropropagação no tempo (*backpropagation through time*), o qual geralmente sofre o efeito da explosão ou perda do gradiente do erro a medida em que dados atuais dependem de valores defasados em um passado distante. Isso é chamado dependência de longo prazo e torna a tarefa de aprendizado da rede muito custosa ou impraticável computacionalmente.

As redes neurais *LSTM*, contudo, não sofrem desse problema uma vez que elas implementam mecanismos de portas (*gates*) capazes de descartar, manter, adicionar ou atualizar informações no tempo de modo a melhor prever o próximo estado e evitar mudanças bruscas da memória o que poderia acarretar a explosão ou perda do gradiente do erro.

Essa característica das redes neurais *LSTM* torna elas ideais para o processamento de dados sequenciais no tempo tais como linguagem natural e tradução. Esse tipo de dado caracteriza-se pelo fato de que a previsão de um estado seguinte depender do estado atual ou de um estado da rede em um momento anterior. Por isso é fundamental que o modelo de rede neural seja capaz de associar de forma eficiente dados atuais a dados remotos no tempo sem os inconvenientes da perda ou explosão do gradiente do erro.

# 3. METODOLOGIA

Esta seção descreve as cinco etapas da metodologia modelada responsável pela geração da estratégia de negociação proposta para aplicações no mercado financeiro: Coleta de Dados, Transformação de Dados, Classificação, Estratégia de Operação e Análise de Resultados, conforme Figura 1.

## 3.1 Base de Dados

As bases de dados utilizadas neste trabalho foram coletadas de forma manual via *download* do sitio da Bovespa[1] de arquivos de cotações históricas contendo atributos relativos ao código do ativo, preços (fechamento, abertura, máximo, mínimo, fechamento anterior), quantidade de negócios, quantidade de papéis, volume financeiro, data e hora das transações realizadas de todas as ações negociadas na bolsa de valores brasileira.

## 3.2 Transformação de Dados

A etapa de transformação dos dados calcula o atributo alvo (*ClasseY*) e realiza a normalização *Log Clusterizado* e *Z-Score* de acordo com o algoritmo utilizado para todos os dados da série temporal de preço original de cotações históricas da Bovespa.

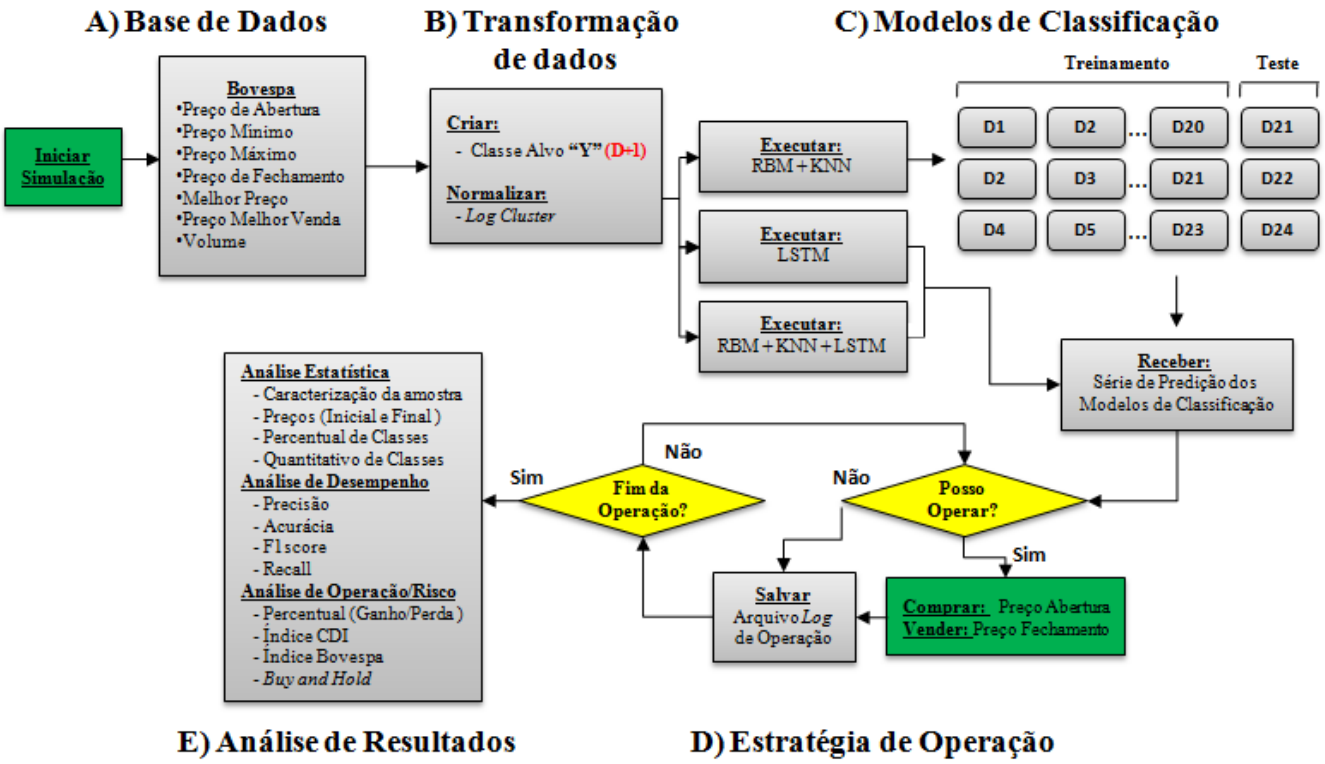
A *ClasseY* é calculada através da diferença entre o preços (fechamento e abertura) do próximo dia. Caso este resultado seja maior do que 0 a variável *ClasseY* recebe o valor de "1". Caso contrário, a variável recebe o valor de "0".

Devido as séries estarem sujeitas a uma quantidade considerável de ruído e variação aleatória, o que pode dificultar a tarefa de previsão, para cada atributo coletado dos dados da Bovespa os dados utilizados nas fases de treinamento e teste foram normalizados de duas formas:

- Algoritmos *KNN* e *RBM*: quando uma série temporal de preço (X) é muito grande, a sua média ($\mu$) e o seu desvio padrão ($\sigma$)

[1]Bovespa: *www.bmfbovespa.com.br*

**Figure 1: Metodologia**




podem não representar a realidade desta mesma série temporal. Desta forma, propõe-se *clusterizar* esta mesma série temporal em $\alpha$ partes para que estas várias amostras de ($\mu$) e ($\sigma$) possam representar a série em questão avaliada, sendo detalhado o cálculo na Equação 10.

$$LogClusterizado = \frac{\chi - (\sum_{j=i}^{\alpha} \mu)}{\sum_{j=i}^{\alpha} \sigma} \quad (10)$$

- Algoritmo *LSTM*: a primeira camada executa uma normalização por *batch* – *BatchNormalation*, onde para cada *batch* utilizado no treinamento a rede executa uma normalização com base nas colunas dos atributos das entradas. Essa normalização é do tipo *Z-Score* ($Z$), onde a diferença cada intervalo ($v'$) com o valor a média ($\mu$) será dividida pelo desvio padrão ($\sigma$) da série avaliada, uma vez que mantém a média dos valores das entradas próximo de 0 e o desvio padrão próximo de 1, conforme Equação 11.

$$Z = \frac{v' - \mu}{\sigma} \quad (11)$$

## 3.3 Modelo de Classificação

Para a etapa de classificação foram propostos três modelos de previsão combinados utilizando os algoritmos *KNN*, *RBM* e *LSTM*. Inicialmente propomos a combinação do RBM com a finalidade de reduzir a dimensionalidade dos dados produzindo a entrada combinada para a utilização do *KNN*. Em seguida avaliamos o algoritmo *LSTM* de forma isolada. E por último, propomos um modelo híbrido entre o (*RBM + KNN + LSTM*) com o intuito de avaliar de possibilidade de melhoria das medidas de desempenho destes modelos propostos.

Durante as fases de calibragem e validação do modelo (*RBM + KNN*), será utilizada uma janela deslizante de 20 dias (Treinamento) e 1 dia (Teste) para verificar toda a série temporal dos 3 códigos de ativos avaliados. Já para o *LSTM* será utilizada 2 anos (2013 a 2014), em torno de 486 dias úteis, na fase de (Treinamento) para a verificação deste algoritmo em 1 ano (2015), ou seja, 243 dias úteis na fase de (Teste).

Como resultados dos 3 modelos de classificação propostos, será gerado uma nova série temporal contendo os sinais de saída da predição (0=caiu 1=subiu) como indicador de tendência dos períodos avaliados.

## 3.4 Estratégia de Operação

O Algoritmo 1 ao identificar uma oportunidade de negociação, realizará ordens de compra e venda (gatilhos) diários para serem iniciados e finalizados no mesmo dia. Caso o sinal da predição da série temporal avaliada gerada pelo modelo de classificação 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 predição seja igual a "0", o gatilho será cancelado.

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

## 3.5 Analises e Resultados

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

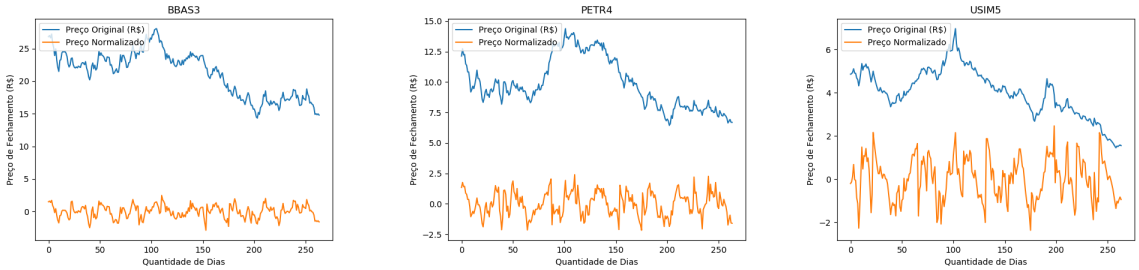



**Figure 2: Série de Preço de Fechamento Original e Normalizada de 243 dias úteis.**

**Algoritmo 1:** ESTRATÉGIA DE OPERAÇÃO

**Entrada:** $SinalPredicao$, $QtdeDiasPredicao$
**Saída:** Arquivo $LogOperacao$ do modelo de classificação.
**início**
  **para** $j$ < $QtdeDiasPredicao[i]$ **faça**
    **if** $SinalPredicao[j]$ = *1* **then**
      comprar ($PreoAbertura(D+1)$);
      vender ($PreoFechamento(D+1)$) ;
      calcular ($LogOperacao[i,j]$);
    **end**
  **fim**
  gravar ($LogOperacao[i,j]$)
**fim**
**retorna** $LogOperacao$

- Análise Estatística: apresenta a caracterização da amostra de cada código de ativo avaliado 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.

- Análise de Desempenho: avaliação dos modelos de aprendizado de máquina através das métricas previamente descritas (acurácia Eq.5, *recall* Eq.6, precisão Eq.7 e especificidade Eq.9) com a finalidade de medir as taxas de desempenho dos classificadores durante a simulação realizada.

- Análise de Operação / Risco: apresenta os resultados dos para cada os código de ativo avaliados, comparando os resultados dos Percentuais (%) e quantitativos acumulados (Perda, Ganho e Retorno Total), o *Buy and Hold*[2], o Acerto total[3] e os Indicadores de mercado (CDI[4], Poupança, Ibovespa[5], Volatilidade[6] e índice *Sharpe*[7]).

[2]*Buy and Hold*: comprando no início da série e vendendo no final do período avaliado.
[3]Acerto total: é o máximo de ganho, caso ocorra o acerto de todos os gatilhos de ganhos.
[4]Certificados de Depósitos Interbancários (CDI): é uma média dos juros praticados entre os Bancos, e serve como uma referência para o preço do dinheiro na economia, pois é utilizada como *benchmark* em muitos investimentos
[5]Ibovespa é o índice da Bolsa de Valores de São Paulo, sendo um indicador de desempenho das ações negociadas na Bovespa
[6]Volatilidade: indicator mede a diferença entre os preços máximos e mínimos, utilizada para indicar os topos e fundos do mercado cambial
[7]índice *Sharpe*: avalia a rentabilidade e o risco de um investimento, sendo fundamental para mesurar o quanto de retorno excedente em relação a um ativo livre de risco é compensado através de seu nível de risco.

A metodologia proposta avalia a viabilidade, o grau de risco de investimento das propostas de estratégias de negociação utilizando técnicas de *machine learning* e *deep learning* e, a possibilidade de aplicação futura destas estratégias em um cenário real no mercado financeiro.

# 4. SIMULAÇÃO E RESULTADOS

A simulação realizada utilizou um arcabouço contendo todas as etapas da metodologia proposta na Seção 3, implementado em linguagem de programação *Python*, acessando os recursos e funções da biblioteca *Scikit-learning* para aplicações com algoritmos de aprendizado de máquina.

A avaliação contemplou um período de 243 dias úteis referentes ao ano de 2015, verificando 03 códigos de ativos (i.e., BBAS3, PETR4, USIM5) 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, bancos e siderurgia, conforme detalhado na Tabela 4.

O experimento realizou a combinação de 3 algoritmos de classificação (*RBM + KNN*), *LSTM* e por último uma combinação híbrida de (*RBM + KNN + LSTM*), avaliando os modelos de forma independente em relação à taxa de precisão. A calibragem dos parâmetros de configuração dos classificadores ocorreu conforme a Tabela 1 após várias simulações realizadas, obtendo com estas configurações, o melhor resultado em relação aos dados utilizados.

A Figura 2 apresenta a série de preço de fechamento em (R$) avaliada para cada código de ativo juntamente com a sua normalização aplicando a técnica logaritmo retorno *clusterizado* detalhado na Equação 10 com a finalidade de retirar *outlier* e transformar o dado bruto para melhorar o desempenho do treinamento do modelo de classificação proposto.

A Tabela 2 apresenta o *baseline* do período avaliado, contendo os indicadores de mercado (CDI, Selic, IGP-M e o Ibovespa). Este último apresentou valor negativo em relação ao período anualizado desta série temporal e também, em relação ao retorno financeiro do *Buy and Hold* dos 03 códigos de ativos avaliados, conforme a Tabela 3.

Os resultados apresentados contemplam a descrição dos atributos da séries temporais avaliadas, o percentual de distribuição das classes, o oráculo [8], o percentual de retorno financeiro do *buy and hold*, a quantidade de gatilhos (perda, ganho e total), o percentual de retorno financeiro da estratégia, a precisão do classificador, as

[8]Oráculo: que é o gabarito de todas as classes "1"de alta na série.

**Table 1: Configuração dos algoritmos *LSTM, KNN* e *RBM*.**

| LSTM | KNN | RBM |
|---|---|---|
| Quantidade de Camadas = 4 | Número de Vizinhos = 5 | Número de Componentes = 4 |
| Camada 1 = Batch Normalization | Tamanho da Folha = 30 | Número de Épocas = 200 |
| Camada 2 = 125 unidades | Ponderação de Pesos = Uniforme | Taxa de Aprendizado = 0,01 |
| Camada 3 = 75 unidades | Distância entre pontos = Euclidiana | Inicialização = Determinístico |
| Regularizador L1 e L2 = 0,005 | | |
| Função de Ativação = Sigmóide | | |
| Otimizador = Adagrad | | |
| Função de Perda = Binary Cross Entropy | | |
| Número de Épocas = 200 | | |
| Tamanho do Batch = 100 | | |
| Conjunto de Validação = 20 | | |

**Table 2: Indicadores Comparativos de *Baseline***

| | |
|---|---|
| **Período Inicial** | 05/01/2015 |
| **Período Final** | 28/12/2015 |
| **Dias úteis de avaliação** | 243 |
| **Taxa do CDI** | 12,5% |
| **Taxa SELIC** | 12,52% |
| **Índice IBOVESPA** | -12,32% |
| **Índice IGPM** | 10,08% |
| **Custo Operacional** | 25,00% |

medidas de risco índices (*Sharpe* e *Sortino*), dentre outros indicadores. E na Tabela 3 os valores destacados em vermelho são as perdas acumuladas em percentual gerados pela estratégia de operação, e/ou alguns *buy and hold* que foram negativos nos períodos avaliados. Já os valores destacados em verde são os percentuais de retornos financeiros gerados pela estratégia de operação e o percentual de precisão, ou seja, a taxa de acerto gerada pela modelo de classificação proposto.

Foi considerado um custo operacional de 25% (imposto de renda *daytrade* de 20% e 5% de custo de transação) em relação ao ganho bruto encontrado pela estratégia de negociação proposta durante a fase de operação. Em seguida, consolidou-se os resultados obtidos na Tabela 3, na qual apresenta o ganho líquido em percentual (%) dos 3 códigos de ativos avaliados, com destaque na coloração verde para os percentuais positivos contendo valores acima dos indicadores financeiros de mercado, conforme a Tabela 2.

Para facilitar o entendimento dos indicadores gerados como resultados nas simulações realizadas, foram consolidados por código de ativo os retornos das análises estatísticas, as medidas de desempenho, as medidas de risco/retorno das séries avaliadas.

A Tabela 3 apresenta os resultados simulados para um período de 243 dias úteis, onde as classes (classe0 e classe1 Real (%)) para os 3 códigos de ativo estavam desbalanceadas, ou seja, a distribuição das classes avaliadas não estavam próximas à 50%, contendo uma maior quantidade de perdas e consequentemente um *buy and hold* negativo. O *buy and hold* encontrado foi negativo para os 3 códigos de ativos, com perdas variando de -66,74% a -26,70%. Para os modelos de classificação híbridos (*RBM + KNN*) e (*RBM + KNN + LSTM*) todos os Percentuais de Retornos financeiros líquidos encontrados foram valores acima dos indicadores de *buy and hold* variando entre -6,09% e 36,98%.

A Precisão do modelo de classificação (*RBM + KNN*) ficou entre 46,25% e 56,51%, e mesmo em uma série temporal com contendo uma tendência de queda com dados desbalanceados, o modelo proposto gerou um lucro líquido para os códigos de ativos PETR4, USIM5 e BBAS3 com valores positivos e ainda acima de qualquer aplicação de renda fixa conforme Tabela 2.

Já para modelo de classificação (*LSTM*) a precisão ficou entre 41,00$ e 46,00% e o Percentual de retorno Líquido de -60,86% e -0,54%, apresentando um desempenho muito péssimo na previsão e com valores negativos na fase de operação, onde somente para os códigos de ativo PETR4 e BBAS3 os valores ficaram abaixo do que o indicador de *buy and hold* conforme Tabela 2.

Por último, o modelo de classificação híbrido (*RBM + KNN + LSTM*) em relação à taxa de precisão variou entre 44,63% e 51,09% e o Percentual de Lucro Líquido com 40,28% e -6,09%, apresentando o melhor retorno financeiro dos modelos avaliados para o código de ativo PETR4.

Em relação aos índices (*Sharpe* e *Sortino*) para os 3 códigos de ativos avaliados sempre apresentaram valores negativos quando o Percentual de Retorno Financeiro Líquido e Bruto foram negativos, confirmando desta forma, a tendência da série em relação ao risco retorno esperado.

Com os resultados apresentados na Tabela 3 fica evidente que nem sempre uma taxa de precisão acima de 50% de acerto na etapa de classificação é sinônimo de ganho financeiro na etapa de estratégia de operação. Um exemplo disso, é o modelo de classificação (*RBM + KNN*) com uma taxa de precisão de 56,18% gerar uma retorno financeiro líquido de 25,66% e o modelo de classificação (*RBM + KNN + LSTM*) com taxa de precisão de 51,09% gerar um retorno financeiro líquido de 32,22% para o código de ativo PETR4.

## 5. CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho foram implementados 3 modelos de classificação compostos pelos algoritmos de (*RBM + KNN*), *LSTM* e (*RBM + KNN + LSTM*), com a finalidade de realizar a seleção e extração de características dos atributos envolvidos e, consequentemente melhorar o aprendizado dos modelos no quesito precisão.

Em seguida a série temporal contendo as previsões de tendencias dos modelos de classificação propostos foram aplicadas como estratégias de negociação contemplando algumas das principais variáveis que envolve o mercado financeiro, como por exemplo: custo operacional, imposto de renda, indicadores de risco/retorno, dentre outros.

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

**Table 3: Resultados de 243 dias úteis do período de 05/01/2015 a 31/12/2015.**

| Modelos | RBM + KNN | | | LSTM | | | RBM + KNN + LSTM | | |
|---|---|---|---|---|---|---|---|---|---|
| **Ativos** | **PETR4** | **USIM5** | **BBAS3** | **PETR4** | **USIM5** | **BBAS3** | **PETR4** | **USIM5** | **BBAS3** |
| **Ganho Máximo (%)** | 10,99 | 9,6 | 7,61 | 8,04 | 8,45 | 7,61 | 5,34 | 8,45 | 7,61 |
| **Perda Máxima (%)** | -7,53 | -10,94 | -5,29 | -6,94 | -15,35 | -6,81 | -5,95 | -10,94 | -5,02 |
| **Perda (%)** | -96,42 | -122,14 | -78,27 | -111,59 | -138,7 | -154,59 | -12,25 | -58,92 | -46,01 |
| **Ganho (%)** | 122,09 | 147,28 | 107,85 | 111,05 | 90,01 | 126,84 | 44,47 | 54,06 | 57,94 |
| **Qtde Perda** | 39 | 38 | 43 | 51 | 43 | 65 | 6 | 16 | 23 |
| **Qtde Ganho** | 50 | 42 | 37 | 43 | 30 | 49 | 17 | 15 | 18 |
| **Qtde Total** | 89 | 80 | 80 | 94 | 73 | 114 | 23 | 31 | 41 |
| **Qtde Classe0** | 153 | 162 | 162 | 148 | 169 | 128 | 219 | 211 | 201 |
| **Qtde Classe1** | 89 | 80 | 80 | 94 | 73 | 114 | 23 | 31 | 41 |
| **Lucro Líquido (%)** | **25,66** | **25,14** | **29,58** | **-0,54** | **-48,69** | **-27,75** | **32,22** | **-4,87** | **11,93** |
| **Lucro Bruto (%)** | **32,08** | **31,43** | **36,98** | **-0,68** | **-60,86** | **-34,69** | **40,28** | **-6,09** | **14,91** |
| **Precisão (%)** | **56,18** | **52,50** | **46,25** | **46,00** | **41,00** | **43,00** | **51,09** | **46,75** | **44,63** |
| **Acuracia (%)** | **60,33** | **63,64** | **54,96** | **53,00** | **56,00** | **51,00** | **56,67** | **59,82** | **52,98** |
| **F1score (%)** | 51,02 | 48,84 | 40,44 | 43,00 | 36,00 | 45,00 | 47,01 | 42,42 | 42,72 |
| **Recall** | 46,73 | 45,65 | 35,92 | 40,00 | 33,00 | 48,00 | 43,37 | 39,33 | 41,96 |
| **Especificidade** | 71,11 | 74,67 | 69,06 | 62,00 | 71,00 | 53,00 | 66,56 | 72,84 | 61,03 |
| **Índice Sharpe** | 2,48 | 4,46 | 1,87 | -0,49 | -19,31 | -3,43 | 6,16 | -2,58 | 1,16 |
| **Índice Sortino** | 4,18 | 6,94 | 3,48 | -0,83 | -30,03 | -6,39 | 10,38 | -4,01 | 2,17 |

**Table 4: Valor de Mercado e Variação dos Códigos de Ativos entre Janeiro de 2015 e 2018.**

| Ativo | Nome | Setor | 2015 | 2018 | Variação |
|---|---|---|---|---|---|
| BBSA3 | Banco do Brasil | Financeiro | R$ 22,58 | R$ 32,93 | 45,84% |
| USIM5 | Usinas Siderúrgicas SA | Mineração | R$ 4,75 | R$ 10,03 | 111,16% |
| PETR4 | Petróleo Brasileiro SA | Óleo e Gás | R$ 9,14 | R$ 16,55 | 81,07% |

**Table 5: Caracterização da Amostra**

| Ativo | BBAS3 | PETR4 | USIM5 |
|---|---|---|---|
| **Preço Inicial** | R$ 22,58 | R$ 9,14 | R$ 4,75 |
| **Preço Final** | R$ 14,98 | R$ 6,70 | R$ 1,58 |
| **Classe0 Real (%)** | 57,61% | 55,97% | 62,14% |
| **Classe1 Real (%)** | 42,39% | 44,03% | 37,86% |
| **Qtde Classe0 Real** | 140 | 136 | 151 |
| **Qtde Classe1 Real** | 103 | 107 | 92 |
| **BuyAndHold (%)** | **-33,66%** | **-26,70%** | **-66,74%** |

cro.

Não é trivial replicar e calibrar todas as variáveis que envolvem e influenciam o mercado de capitais em especial a Bolsa de Valores, no entanto, através do arcabouço metodológico proposto, verifica-se, que muitos indicadores ou os principais de análise estatística, desempenho e de risco utilizados por especialistas de mercado, foram implementados e consolidado em um único ambiente.

Os resultados apresentados na Tabela 3 obtidos pelo modelo classificação (*RBM + KNN*) avaliando os 3 códigos de ativos de setores diversos da economia, acrescido de um custo operacional de 25% apresentaram valores acima dos principais indicadores (*baselines*) de referência conforme Tabela 2, o que mostra a total viabilidade de aplicarmos o modelo de estratégia proposta em um cenário real no mercado financeiro.

O modelo de classificação *LSTM* de forma isolada apresentou somente valores de percentual financeiro líquido negativos. E este resultado é justificado devido a taxa de precisão deste classificador não ter atingido valores acima de 50,00% conforme 3, se posicionando desta forma, como um modelo aleatório, que não possui uma taxa mínima de confiança.

Por fim, o modelo de classificação híbrido (*RBM + KNN + LSTM*) encontrou o melhor percentual de retorno líquido para o código de ativo PETR4, sendo superior aos demais modelos propostos. Porém, para os demais códigos de ativos BBAS3 e USIM5, este modelo apresentou valores negativos inviabilizando a possível aplicação deste modelo como estratégia de operação em uma ambiente real do mercado financeiro.

Desta forma, o tema deste artigo envolve uma área interdisciplinar que abre oportunidades para o desenvolvimento de trabalhos futuros envolvendo diversas áreas do conhecimento, contendo vários desafios de pesquisa. Almeja-se realizar simulações / *backtesting* destes modelos propostos de classificação (*RBM + KNN*) e (*RBM + KNN + LSTM*) validando-os através de simuladores realísticos com a possibilidade de validação real no mercado financeiro contemplando todas variáveis que envolve todo esse ambiente operacional.

## Referências


[1] G. E. Abdulaziz Almalaq. A review of deep learning methods applied on load forecasting. *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[2] I. Aldridge. *High-frequency trading: a practical guide to algorithmic strategies and trading systems*. John Wiley & Sons, 2013.

[3] Y. Bengio. *Learning Deep Architectures for AI*. Foundations and Trends in Machine Learning, 2009.

[4] K. P. V. George S. Atsalakis. Surveying stock market forecasting techniques – part ii: Soft computing methods. *Expert Systems with Applications*, 2009.

[5] E. . Y. Global. Find the right people, processes and technology to manage record-to-report risks. *Managing*

*Operational Tax Risk*, 2014.

[6] A. Grimes. *The Art and Science of Technical Analysis: Market Structure, Price Action, and Trading Strategies*, volume 1. Editora Saraiva, 2012.

[7] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[8] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[9] S. J. Hochreiter S. Long short-term memory. *Neural computation*, 9:1735–1780, 1997.

[10] R. R. Kevin Beyer, Jonathan Goldstein. When is "nearest neighbor"meaningful? *CiteSeerX*, 1999.

[11] A. V. Lemos. *Comissão de Valores Mobiliários (CVM) - Programa de Treinamento de Professores*. Análise Técnica (Teoria de Dow, Elliott, Gráficos e Indicadores Técnicos), 2013.

[12] Z. Z. M Zhang. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2014.

[13] D. S. Mansooreh Kazemilaria, Abbas Mardanib. An overview of renewable energy companies in stock exchange: Evidence from minimal spanning tree approach. *Renewable Energy*, 102:107–117, 2017.

[14] A. F. B. e. L. L. Martín Iglesias Caride. Stock returns forecast: An examination by means of artificial neural networks. *Complex Systems: Solutions and Challenges in Economics, Management and Engineering*, 2017.

[15] T. Mitchell. *Machine Learning*, volume 1. McGraw-Hill Science/Engineering/Math, 1997.

[16] S. Niaki and S. Hoseinzade. Forecasting s&p 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, 9(1):1–9, 2013.

[17] R. J. S. Paulo Henrique Kaupa. Rough sets: technical computer intelligence applied to financial market. *International Journal of Business Innovation and Research*, 13, 2017.

[18] H. V. Roberts. Stock market patterns and financial analysis: Methodological suggestions. *The Journal of Finance*, 14, 1959.

[19] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Englewood Cliffs, 1995.

[20] F. B. e. J. M. Thomas Lopes. Mineração de opiniões aplicada a análise de investimentos. *Sociedade Brasileira de Computação*, 2008.

[21] A. S. S. Tiago P. Oliveira, Jamil S. Barbar. Computer network traffic prediction: A comparison between traditional and deep learning neural networks. *International Journal Big Data Intelligence*, 3(1), 2016.

[22] G. Varga. índice de sharpe e outros indicadores de performance aplicados a fundos de ações brasileiros. *Revista de Administração Contemporanea*, 2001.

[23] H. White. Economic prediction using neural networks: The case of ibm daily stock returns. In *Neural Networks, 1988., IEEE International Conference on*, pages 451–458. IEEE, 1988.