UNIVERSIDADE FEDERAL DA PARAÍBA (UFPB)
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS (CCSA)
DEPARTAMENTO DE FINANÇAS E CONTABILIDADE (DFC)
CURSO DE CIÊNCIAS ATUARIAIS (CCA)

IGOR LUÍS ALBUQUERQUE SILVA

GOOGLE TRENDS E COMBINAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DO IBOVESPA

JOÃO PESSOA, PB

2020

IGOR LUÍS ALBUQUERQUE SILVA

GOOGLE TRENDS E COMBINAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DO IBOVESPA

Monografia apresentada ao Curso de Ciências Atuariais, do Centro de Ciências Sociais Aplicadas, da Universidade Federal da Paraíba como requisito para a obtenção do grau de Bacharel em Ciências Atuariais.

Orientador: Prof. Me. Filipe Coelho de Lima Duarte

Coorientador: Prof. Dr. Ronei Marcos de Moraes

JOÃO PESSOA, PB

2020

Catalogação na publicação
Seção de Catalogação e Classificação

S586g Silva, Igor Luis Albuquerque.
Google Trends e Modelos de Aprendizagem de Máquina para Predição do Ibovespa / Igor Luis Albuquerque Silva. - João Pessoa, 2020.
47f.

Orientação: Filipe Coelho de Lima Duarte.
Coorientação: Ronei Marcos de Moraes.
Monografia (Graduação) - UFPB/CCSA.

1. Aprendizagem de máquina. 2. Mercado financeiro. 3. Google Trends. I. Duarte, Filipe Coelho de Lima. II. de Moraes, Ronei Marcos. III. Título.

UFPB/CCSA

IGOR LUÍS ALBUQUERQUE SILVA

GOOGLE TRENDS E COMBINAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DO IBOVESPA

Trabalho de Conclusão de Curso para o curso de Ciências Atuariais na UFPB, como requisito parcial à obtenção do título de bacharel em Ciências Atuariais.

BANCA EXAMINADORA

Me. Filipe Coelho de Lima Duarte

Orientador

UFPB

Dr. Ronei Marcos de Moares

Coorientador

UFPB

Dr. Luiz Carlos Santos Junior

Membro Avaliador

UFPB

AGRADECIMENTOS

Em primeiro lugar, eu gostaria de agradecer aos meus pais, Cláudia Regina e André Luís, pelo carinho e apoio que me deram e continuam me dando depois desses anos, sem vocês eu não teria chegado aqui.

Gostaria também de agradecer aos meus orientadores, ao professor Filipe Duarte por me incentivar a programar, sem ele este trabalho não seria possível, e também ao professor Ronei Marcos, pela ajuda em questões metodológicas, e por me incentivar a fazer o mestrado em Estatística.

Agradeço também à minha namorada Carla Nayara, pelo apoio e carinho durante esse processo de escrita do trabalho, pelos feriados separados e por estar sempre presente ao meu lado.

Antes de ingressar no curso, tive e tenho inúmeras amizades desde à época do colégio, dentre as demais, Mário Praun e Rodrigo Santa Cruz se destacam, pelo companheirismo e apoio durante toda minha graduação, mas também, construí boas amizades durante o curso que pretendo levar para o resto da vida, com Horácio Ramalho, pelas conversas filosóficas e planos de empreendedorismo, Karys Emanuelle pelas risadas e brincadeiras entre os intervalos das aulas, Leonarrrrdo pela amizade sincera e conselhos em relação à tudo e por fim, Paula Bianca, pelas inúmeras conversas sobre investimentos e risadas sobre os sardinhas da Bolsa.

Por fim, mas não menos importante, agradeço aos servidores e professores dos departamentos de Finanças e Contabilidade, pelo apoio na coordenação, processos acadêmicos, e, pelo conhecimento que me foi dado durante estes anos.

“Step after step we try controlling our fate

When we finally start living it's become too late

Trapped inside this octavarium”

Dream Theater – Octavarium (2005)

Resumo

O presente estudo teve como objetivo prever o sentido do retorno do Ibovespa a partir dos retornos de nove papeis negociados na B3 durante o período de 2012 até 2018 e a partir do índice Google Trends. O período de 2012 A 2017 foi utilizado como período de treino, enquanto o 2018, como período de teste. Numa primeira etapa, para classificar o sentido do retorno das empresas, utilizou-se de: regressão logística, *KNN, SVM* linear e radial, *Bagging, Adaboost, XgBoost,* e *Random Forest*. Para encontrar os melhores parâmetros para os modelos, foi utilizada a função *grid_search* () do pacote *caret, (Kuhn; 2008),* onde ocorre a validação cruzada para os modelos durante a fase de treinamento. Numa segunda etapa, o Ibovespa foi treinado por uma rede neural com 9 neurônios e *decay* de 0,91. Como os resultados da modelagem inicial não foram satisfatórios, pois apresentaram uma acurácia média de 50%. Assim, optou-se por não utilizar o sentido de retorno das empresas para a predição do Ibovespa, calculando então, ao final, com base na metodologia proposta por este trabalho, o valor de 1,34%, enquanto a estratégia clássica de *Buy and Hold* apresentou o valor de 1,10%. Uma das limitações do estudo é que, só foram utilizadas duas classes para a predição, “SUBIR” e “DESCER”, porém, como sugestão futura, sugere-se a inclusão de uma terceira classe “NEUTRO”.

Palavras-chave: Aprendizagem de máquina; mercado financeiro; *Google Trends*.

Abstract

The present study aimed to show *Google Trends* with machine learning, an end of forecast or Ibovespa return, based on the returns of nine papers traded on B3 during the period from 2012 to 2018. Between 2012 and 2017 it was selected as the period training, while the test was carried out in 2018, using the following models to classify the return of companies: regression logistics, KNN, linear and radial SVM, *Bagging, Adaboost, XgBoost and Random Forest.* To find the best patterns for the models, the *grid_search* () function of the *caret* package (Kuhn; 2008) was used, where cross-validation occurs for the models during the training phase. As for Ibovespa, it was trained by a neural network with 9 neurons and a 0.91 decay. The results of the initial modeling were not satisfactory, with an average accuracy of 50% and they chose not to use the companies' models to predict the market index, calculating, at the end, or the accumulated return of the strategy for making the decision of investments based on the methodology proposed by this work was 1.34%, while the classic purchase and retention strategy was 1.10%. One of the study's permissions is that, therefore, two classes were used for prediction, "UP" and "DOWN", however, as a future suggestion, it suggests an inclusion of the third class "NEUTRAL".

Keywords: Machine learning; financial market; *Google trends.*

LISTA DE SIGLAS

ANN = Artificial Neural Networks

B3 = Bolsa Brasil Balcão

CAPM = Capital Asset Pricing Model

CDI = Certificado de Depósito Interbancário

D-CAPM = Downside Capital Asset Pricing Model

DT = Decision Tree

HME = Hipótese do Mercado Eficiente

NN = Neural Networks

RMSE = Raiz quadrada do erro-médio

SVM = Support Vector Machines

LISTA DE GRÁFICOS

Gráfico 1: Estrutura da rede neural 31
Gráfico 2: Contabilização do sentido do retorno 35
Gráfico 3: Frequência da variável Google Trends 36
Gráfico 4: Preço dos fechamentos dos ativos durante o período 36
Gráfico 5: Fechamento do Ibovespa 37
Gráfico 6: Retorno dos ativos 38
Gráfico 7: Retorno Acumulado Buy and Hold VS Estratégias 40

SUMÁRIO

1 INTRODUÇÃO 14

1.1 Problema de pesquisa 15

1.2 Objetivos 15

1.2.1. Objetivo geral 15

1.2.2 Objetivos específicos 16

1.3 JUSTIFICATIVA 16

2 REVISÃO DE LITERATURA 17

2.1 PRECIFICAÇÃO DE ATIVOS 17

2.2 GOOGLE TRENDS 19

2.3 APRENDIZAGEM DE MÁQUINA 21

3 METODOLOGIA 24

3.1 POPULAÇÃO E AMOSTRA 24

3.2 COLETA E TRATAMENTO DOS DADOS 24

3.3 Análise dos dados 25

3.3.1 Florestas aleatórias 26

3.3.2 Regressão Logística 26

3.3.3 Bagging 27

3.3.4 Boosting 27

3.3.5 Máquinas de Vetores de Suporte 28

3.3.6 K- vizinhos próximos 28

3.3.7 XGBoost 29

3.4 Avaliação de modelos 30

3. 5. Redes neurais artificiais 31

3.6 Comparação de estratégias de investimento 32

4 RESULTADOS 33

4.1 Estatísticas descritivas 33

4. 2. Modelagem preditiva 38

5 CONCLUSÃO 42

REFERÊNCIAS 43

1 INTRODUÇÃO

O mercado de capitais é formado por Agentes deficitários e superavitários, os quais buscam captar recursos a fim de melhorar sua atividade operacional em troca de ativos financeiros, debêntures (títulos empresariais), por exemplo, em troca de menos risco, e ações (porcentagem do patrimônio líquido da empresa), em troca de um maior risco, que pode ser tido como a volatilidade do mercado.

Assim, pela volatilidade, os investidores buscam maximizar seus retornos. Contudo, também é o sistema que pode gerar a maior perda caso a estratégia não seja bem definida. Como pode ser visto em Markowitz (1952), que introduziu à análise de investimentos uma maneira de gerenciar uma carteira de ativos com ferramentas estatísticas como média e desvio-padrão, sendo a primeira, o retorno proporcionado por uma carteira de ativos, e a segunda, o risco proporcionado por uma carteira de ativos. O retorno pode ser entendido como o quanto o investidor espera lucrar em uma determinada aplicação. Já o risco refere-se à probabilidade de o retorno não se realizar, ou seja, dados dois ativos com o mesmo retorno, o investidor tende a escolher aquele com menor risco.

Apesar disso, há investidores profissionais e qualificados no mercado que buscam atingir retornos superiores ao *benchmark*, representado, no Brasil, pelo índice do Ibovespa (Ibov). Para estes investidores, são necessárias informações atualizadas sobre as empresas em que investem. Nesse sentido, há a divulgação dos demonstrativos financeiros no site do órgão regulador (i.e., Comissão de Valores Mobiliários - CVM), de notícias, em plataformas de buscas - como a *Google*, que em 2011 disponibilizou, por meio da ferramenta *Google Trends*, o volume de busca das demais buscas feitas pelos usuários. Essa plataforma da *Google* vem influenciando pesquisadores como Fondeur (2012), Kristoufek (2013), Kristoufek et al. (2016), Kim et al. (2018) a utilizarem a ferramenta como um diferencial em seus estudos.

Com o avanço computacional, métodos algorítmicos e estatísticos vêm sendo implementados com vistas a alcançar a maior maximização dos retornos, como se verifica em Tsai (2009), Patel et al. (2015) e Preet et al. (2018). Como exemplo, têm-se os “robôs” que operam na compra e venda de ações e os modelos que realizam previsões de como o mercado pode se comportar, inclusive na abertura do próximo pregão.

Trabalhos nacionais e internacionais demonstram que é possível, com a modelagem preditiva e através de algoritmos de aprendizagem de máquina, obter retornos maiores que o índice utilizado como *benchmark* e predizer um índice de mercado. Para realizar tais estudos, não se utilizam somente as variáveis financeiras, afinal, o mercado é um ambiente dinâmico onde notícias e informações da *internet* impactam o mercado, de forma positiva, ou negativa. Se destacam, os trabalhos nacionais: Vargas et al. (2017), Silva *et. al* (2019), Marreti *et. al* (2019) e os internacionais: Pagolu et al. *(2016)*, Xiong et al. (2016), Preet *et al.* (2018).

Através dos estudos citados e argumentos demonstrados, cogitou-se a inclusão da ferramenta *Google Trends* com diversos métodos de aprendizagem de máquina, para a formação de uma "carteira teórica" composta por nove ações com maior volume de negociação, tendo objetivo de acompanhar um dos índices de mercado da bolsa de valores brasileira, o Ibovespa.

1.1 Problema de pesquisa

A partir desse panorama, este trabalho pretende responder ao seguinte questionamento: Como o *Google Trends* e os Modelos de Aprendizagem de Máquina podem melhorar as previsões da direção do Ibovespa?

1.2 Objetivos

Este trabalho tem como objetivo responder a objetivos gerais e o específico.

1.2.1. Objetivo geral

O intuito deste trabalho foi prever a direção do sentido do IBOVESPA através de combinação de modelos de aprendizagem de máquina e por uma rede neural com nove ativos negociados na bolsa de valores brasileira (B3), entre 2012 e 2018.

1.2.2 Objetivos específicos

- Utilizar a Aprendizagem de Máquina para prever a direção dos movimentos dos preços das ações selecionadas;
- Combinar as previsões das ações e o *Google Trends* por meio de uma Rede Neural Artificial para prever a direção do Ibovespa;
- Comparar o retorno da estratégia por meio da previsão do Ibovespa com o retorno da estratégia clássica de investimentos (*Buy and Hold*).

1.3 JUSTIFICATIVA

Diante do contexto de interseção global, em que as economias mundiais estão conectadas com a revolução causada pela internet, pessoas físicas passaram a se conectar com mais intensidade, empresários e investidores de diversos mercados acionários globais. Por exemplo, algum gestor de alguma empresa, ao comentar sobre a sociedade em suas redes sociais, pode causar oscilações inesperadas no mercado. Conforme observado por Kim et al. (2018), é possível observar a volatilidade do mercado com as pesquisas do *Google Trends*.

Além do mais, há um crescimento exponencial nos dados produzidos pelos humanos, o que leva à necessidade de modelos estatísticos e computacionais cada vez mais poderosos, capazes de obter resultados rápidos para a tomada de decisão, conforme visto por Patel et al. (2015), que demonstram que, com os algoritmos de aprendizagem de máquina, se diminui o erro de previsão de índices do mercado. Isso foi confirmado por Machado et al. (2018), que através de modelos híbridos de aprendizagem de máquina e aprendizagem profundo, conseguiram desenvolver um algoritmo de compra e venda para operar no mercado acionário brasileiro.

Nesse contexto, o presente estudo se mostra relevante com o fato de que há a junção destes dois pilares descritos anteriormente, tanto na questão computacional, quanto na questão das pesquisas através da internet, já que o campo de estudo que relaciona finanças, e, aprendizagem de máquina, ainda está em crescimento em âmbito nacional.

2 REVISÃO DE LITERATURA

Nesta seção, foram levantados estudos empíricos dos tópicos relevantes para a construção teórica deste trabalho. O primeiro tópico, *Precificação de ativos*, demonstra a construção dos primeiros modelos de mensuração do retorno de algum ativo, o *CAPM*, passando pela teoria da Hipótese do Mercado Eficiente. O segundo tópico, *Google Trends*, demonstra estudos sobre a variável que foi empregada no estudo, demonstrando que, com a inclusão dela, os pesquisadores obtiveram bons resultados, em comparação com o não uso. E, por fim, o último tópico, *Aprendizagem de máquina*, compila estudos que utilizaram as demais técnicas utilizadas por esta monografia.

2.1 PRECIFICAÇÃO DE ATIVOS

Lintner (1964) e Sharpe (1966) deram início ao modelo *CAPM* de duas fases, um modelo de precificação de ativos, a fim de calcular o retorno que teria com algum ativo financeiro, dado uma taxa livre de risco, e o retorno do mercado. Mais tarde, Fama e French (1993) introduziram, para a precificação das ações, o modelo do *CAPM* contendo três fatores relacionados com o *book-to-market* – índice que mede a oportunidade de crescimento das empresas - com o tamanho da empresa. No artigo, introduziram também fatores para a precificação de títulos, porém, para a presente monografia, não serão discutidas questões dos títulos, dado que o principal objeto deste trabalho é o mercado acionário brasileiro.

O precursor na área de precificação de ativos foi Sharpe (1964), que buscou estender a equação de comportamento do investidor na relação risco x retorno, através de uma equação de equilíbrio derivando funções de utilidade, a fim de mostrar uma zona de "conforto" para o investidor racional em relação a ativos de risco, averiguando que, para investidores racionais, tende-se a aceitar baixas expectativas de retornos, aceitando o baixo risco. Lintner (1965) analisou, de forma matemática, problemas relacionados na seleção de ativos de risco e, embora esteja sobre condições idealizadas, concluiu que o retorno mínimo que o investidor deve ter é algo denominado como risco de capital.

Mais tarde, Sharpe (1966) buscou medir o risco de fundos de investimentos americanos utilizando o retorno de um ativo, uma taxa livre de risco e o risco do determinado ativo. Como resultado da pesquisa, criou um índice que mede a performance de qualquer ativo de renda variável.

A partir de outro panorama, Fama (1970) buscou fazer testes a fim de verificar as eficiências do mercado financeiro. Ele testou o mercado sob três formas de eficiência: forma fraca, onde se observa apenas os preços históricos dos ativos; a semiforte, que diz onde os preços se ajustam a informações públicas, ou seja, novas informações já estão precificadas; a forte, que diz respeito aos investidores que possuem informação privilegiada e se baseiam nelas para tomada de decisão na compra ou venda de um ativo. Através de testes estatísticos, identificou que os mercados atuam sobre a forma semiforte, onde novas informações são precificadas pelos acionistas. Em seguida, Fama e French (1993) identificaram cinco fatores de risco nos retornos das ações mediante o uso de regressão linear e a fim de observar quais fatores de risco influenciam cada ativo, relacionados com o tamanho da empresa, valor de mercado, e, para títulos, o período de maturação, e o risco envolvido.

Tendo em vista a Hipótese proposta por Fama, diversos autores buscaram estudar se de fato as bolsas de valores eram eficientes, ou seja, se os preços refletiam toda a informação disponível. Marques (2015), buscou analisar a eficiência do mercado brasileiro com base na teoria das finanças comportamentais e obteve como resultado que os mercados se encontram na forma fraca, contrariando a Hipótese de Mercado Eficiente (HME). Por outro panorama, Caride *et al.* (2017) buscou avaliar a HME no mercado de ações brasileiro através da negociação de alta frequência por meio de redes neurais, certificando-se de que é possível superar o retorno de mercado, indo contra teoria proposta por Fama.

Posteriormente, Paiva (2005) analisou o *CAPM* e o *D-CAPM* e buscou avaliar se o último é uma alternativa eficiente para a precificação de ativos. Coletou os retornos das 40 empresas listada na Bolsa de Valores de São Paulo, de 1996 até 2002. Para utilizar proxies da taxa livre de risco e risco com o Certificado de Depósito Interbancário, e o retorno de mercado, o índice Ibovespa. O autor identificou que o *D-CAPM* tem uma maior capacidade de explicação dos retornos dos ativos que o *CAPM*.

Em seguida, Santos et al. (2011) procuraram incluir mais uma variável no modelo de precificação de ativos, sendo esta o risco no momento. Para calcular a nova variável, utilizaram o retorno acumulado das ações brasileiras, tendo como resultado que o modelo é válido para o mercado acionário brasileiro.

Logo depois, Noda et al. (2016) incluíram mais um fator no modelo de três fatores de Fama e French, o risco lucro/preço, utilizando o índice lucro/preço como indicador *ex ante* para explicar o retorno das empresas brasileiras de 1995 até 2013, concluindo que o fator de risco aplicado ao mercado acionário brasileiro é relevante para o modelo apresentado por Fama e French.

Por fim, Silva et al. (2019), buscaram fazer a *valuation* de uma empresa brasileira com o custo capital sendo simulado probabilisticamente pela técnica de Monte Carlo. Introduziram os riscos de um país emergente na *valuation* para calcular as premissas do custo de capital, utilizando as taxas em uma distribuição normal, apesar de não preverem o preço com exatidão. Mesmo assim, o preço da avaliação determinística e o estocástico foram próximos.

Observando os artigos supracitados, é possível verificar que o campo de aplicação para a modelagem financeira é vasto, incluindo, modelos estocásticos, o que permite a inserção de algoritmos estatísticos e computacionais mais avançados para a predição dos dados das ações.

2.2 GOOGLE TRENDS

O *Google Trends* é uma plataforma de pesquisa da *Google* que coleta os dados e divulga a quantidade de dados que foram pesquisados ao longo de um determinado período. Segundo Choi e Varian (2012), a plataforma providencia os dados dos volumes das pesquisas geográficas em série temporal.

O primeiro estudo através do qual foi observado o uso de buscas na internet para pesquisas científicas foi feito por Michael *et al.* (2005), que estudaram o potencial da Internet para prever a taxa de desemprego nos Estados Unidos. Utilizando o método de regressão linear, em que quiseram determinar se os dados coletados da taxa de desemprego

são influenciados pelos dados das pesquisas diárias sobre busca de emprego na Internet, foi encontrado um sinal positivo significante entre os dados e a taxa de desemprego.

Posteriormente, Carneiro (2009) buscou introduzir a ferramenta *Google Trends* para profissionais de saúde no rastreamento de doenças. Foi criada uma plataforma de busca, a *Google Flu Trends*, onde há dados com informações de alguns vírus transmissores de gripe, incluindo a aviária. Concluiu que o *Google Trends* pode ser melhor ajustado para rastrear doenças epidêmicas rapidamente que os métodos tradicionais.

Choi e Varian (2012) tinham como objetivo disseminar o *Google Trends* entre o público e ilustrar como é possível prever com os dados oriundo da plataforma. Demonstrando exemplos com modelos autorregressivos (AR), por exemplo, do desemprego americano, descobriram que modos com dados do *Google Trends*, superam o poder explicativo de modelos que não utilizam tais dados.

De forma a analisar a plataforma no âmbito de variáveis macroeconômicas, Fondeur e Karamé (2012) buscaram investigar se os dados do *Google* podem melhorar os modelos de previsão da taxa de desemprego entre os jovens na França. Utilizaram um modelo estatístico estimado com uma modificação do filtro Kalman que permite o uso de análise multivariada, e descobriram que, com os dados do *Google*, aprimoraram-se os resultados da predição da taxa de desemprego da população francesa entre 15 e 24 anos.

Ao estudar as variáveis macroeconômicas, Carrière-Swallow e Labbé (2013) analisaram se as pesquisas no *Google* podem sinalizar para agentes reguladores sobre o consumo agregado em um país emergente. Primeiramente, criaram um modelo de série temporal autorregressiva a fim de observar a estrutura dos dados obtidos pelo *Google*; depois, criaram um modelo ARMA, encontrando a caracterização mais forte da série temporal. Então, analisaram o R^2 (Coeficiente de Determinação) e o RMSE (Raiz Quadrática do Erro Médio) a fim de analisar a predição e concluíram que os modelos de previsão podem ser melhorados com os padrões de busca através dos dados do *Google*.

No mercado financeiro, a plataforma foi utilizada por Kristoufek (2013) que propõe uma forma de diversificação na carteira através das pesquisas do *Google Trends*, utilizando o desvio padrão e o índice de Sharpe como formas de medir o risco das carteiras e ações. Concluiu que o *Google Trends* pode ser utilizado com para auxiliar na gestão de risco de uma carteira de investimentos.

Na área da psicologia, Kristoufek *et al.* (2016) analisaram se os dados coletados pelo *Google* Trends podem ajudar a calcular estimativas de suicídio na Inglaterra antes de dados governamentais e buscaram analisar como palavras como “depressão” e “suicídio” estão relacionadas a suicídios. Encontraram que o termo “Depressão” não é relacionado positivamente com suicídios, contudo, o termo “Suicídio”, é positivamente relacionado com a ocorrência de suicídios, em outras palavras, de forma preditiva, observaram que o termo “Depressão”, não é relacionado com suicídios, mas, a busca pelo termo “Suicídio”, sim.

Por fim, ao utilizar a ferramenta para o mercado financeiro Kim *et al.* (2018), buscaram compreender se as pesquisas no *Google* explicam atuais e futuros retornos anormais, volume de negociação e volatilidade das empresas listadas na bolsa de Oslo. Encontraram que os dados das pesquisas do *Google* podem prever volume e volatilidade, mas não retornos.

Conforme observado nos estudos, há uma distribuição geográfica entre os trabalhos que utilizam a plataforma, encontrando que, no geral, os dados obtidos pela plataforma tornaram os modelos utilizados mais significativos em sua capacidade explicativa. Vale salientar que Kristoufek *et al.* (2016), atuando na área da psicologia, relacionou as pesquisas do *Google* com as taxas de suicídio.

2.3 APRENDIZAGEM DE MÁQUINA

Aprendizagem de Máquina é uma área do conhecimento que combina estatística e computação com o intuito de desenvolver algoritmos que aprendem com a experiência. Técnicas diversas podem ser utilizadas afim de alcançar predições, sejam os métodos de aprendizagem supervisionados, não supervisionados, ou, os profundos, que este último, é frequentemente utilizado por meio de redes neurais, que possuem alta capacidade preditiva por meio de funções matemáticas. No campo das redes neurais, McCulloch e Pitts (1943) introduziram a ideia de um neurônio computacional baseado no cérebro humano. Após alguns anos, Rosenblatt (1958) deu início ao *Perceptron*, onde o modelo aprendia conforme

um determinado padrão, aprendizagem supervisionada. Um ano após, Samuel (1959) desenvolveu um algoritmo que jogava damas através de uma função de custo. O algoritmo, após um tempo de treino, conseguiu superar as habilidades do autor do código.

Taylor (1967) criou um algoritmo para reconhecimento facial para tentar reconhecer 10 fotos. Após 250 tentativas, o algoritmo conseguiu aprender a classificar as fotos corretamente. Inspirado em Samuel (1959) e em Griffith (1974), comparou dois modelos novos de aprendizagem de máquina, com o proposto por Samuel, e foi encontrado que os dois novos modelos convergem com decisões feitas por jogadores profissionais, em comparação com o polinômio linear de Samuel.

Fukushima (1980), desenvolveu a primeira rede neural com múltiplas camadas, chamada de *Neocognitron*, em que aprendeu de forma autônoma por padrões de letras e números de forma similar ao globo ocular humano. Hill *et. al.* (1994) buscaram comparar os resultados de trabalhos dos modelos estatísticos tradicionais de previsão com modelos desenvolvidos por redes neurais, onde, encontraram que as previsões feitas por redes neurais são superiores aos modelos de estatística tradicionais.

Posteriormente, Tsai e Wang (2009), que tinham como objetivo criar um modelo que combine *ANN* e *DT* para aumentar a acurácia na previsão do preço das ações, testaram os modelos de aprendizagem de máquina separados e então criaram dois modelos híbridos com *ANN + DT*, e um com *DT + DT*. Concluíram que o algoritmo híbrido da rede neural e da árvore de decisão teve um melhor desempenho, alcançando 77% de acurácia, em comparação com os outros: apenas árvores de decisão, 65%; apenas redes neurais, 59%; e combinação entre duas árvores de decisão, 66%.

Para facilitar o uso de alguns usuários com certos algoritmos de *Machine Learning*, Pedregosa *et al.* (2011) objetivaram desenvolver um pacote para a linguagem de programação *Python* contendo diversos algoritmos implementados de aprendizagem de máquina. Os algoritmos foram implementados para a linguagem de alto nível, com as ferramentas de criação de bibliotecas e a biblioteca contém algoritmos de aprendizagem de máquina para análise estatística.

Em prosseguimento com os estudos de predição ao mercado financeiro, Patel *et al.* (2015) buscaram prever o preço de dois índices do mercado indiano. Assim, dividiram a modelagem em duas fases: na primeira, utiliza-se o modelo de *Support Vector Regression*

(*SVR*); e na segunda, uma modelagem híbrida com rede neural artificial, *random forest*, e, *SVR*. Constataram que o modelo híbrido reduziu o erro de predição dos dados.

De outro modo, Preet *et al.* (2018) incluíram variáveis macroeconômicas na predição do preço das ações, como preço das *comodities* e taxa de câmbio, utilizando os modelos *AdaBoost*, *Gradient Boosting*, *Support Vector Machines*, e *Random Forest*. Os autores encontraram correlação forte e positiva entre a bolsa de Bombai e o índice de ouro, tendo o *AdaBoost* como o modelo com maior percentual de predição.

Observando os estudos anteriores, é possível verificar que, a evolução da capacidade computacional permitiu que técnicas mais complexas fossem desenvolvidas. Desde a criação e aperfeiçoamento de algoritmos para jogos de Damas, aliados com maior poder computacional, os algoritmos vêm sendo capazes de desenvolver predições cada vez mais complexas, como redes neurais profundas, redes nebulosas, e reconhecimento de dados através de sons e imagens.

3 METODOLOGIA

Para alcançar os objetivos da pesquisa, a metodologia foi dividida em duas etapas. A primeira etapa foi composta do treinamento de sete modelos de aprendizagaem de máquina, afim de verificar quais são os melhores por empresa selecionada, já a segunda, foi composta da combinação dos melhores modelos, através de uma rede neural, afim de prever com o Ibovespa.

3.1 POPULAÇÃO E AMOSTRA

A população consistiu em todas as ações negociadas na B3, tendo como período de análise o ano de 2012 até o ano de 2018. A escolha desse período decorreu do fato da implementação da plataforma *Google Trends* ter sido apenas a partir de 2012. Por outro lado, a amostra a ser utilizada foi composta por nove ações com os maiores volumes de negociação da B3 durante o período supracitado.

3.2 COLETA E TRATAMENTO DOS DADOS

Os dados financeiros, expostos no Quadro 1, foram coletados através da plataforma Economática®. Para pesquisar as empresas na plataforma, foram utilizados dois tipos de palavra-chave, o nome da empresa e o *ticker* negociado dela na bolsa de valores ex: “Petrobrás”, “PETR4”.

No que diz respeito à variável alvo da modelagem, na primeira etapa, foi o sentido do retorno de cada empresa; já na etapa final, a variável modelada foi o sentido do retorno do Ibovespa.

Para analisar o sentido do retorno de cada empresa, foi necessário calcular o retorno logaritmo dos preços dos ativos, e então classificá-lo de forma binária, sendo empregado *um* caso o ativo tenha tido retorno positivo e *zero* caso tenha tido retorno negativo ou movimentação neutra. Para poder modelar o sentido diário do retorno, foi necessário criar

um *lag* entre as demais variáveis, assim, o sentido do retorno diário, pode ser previsto com a informação passada.

Para analisar o sentido do retorno do Ibovespa (...).Os dados foram separados numa proporção de 85% para treino, 2012 até 2017, e, 15% para teste, 2018.

Quadro 1: Descrição das variáveis

Variáveis	**Detalhamento das variáveis**	**Estrutura das variáveis**
Sentido_RET	Variável binária que demonstra o sentido do retorno por empresa	Categórica (1 – "Subir"; 2 – "Descer ou neutro")
RET	Retorno logarítmico das empresas	Numérica
VOLUME	Volume de negociação das empresas da Bolsa até fechamento do pregão	Numérica
Close	Preço de fechamento no pregão anterior	Numérica
Open	Preço de abertura no pregão anterior	Numérica
DIFF	Mede a diferença entre o preço de fechamento com o preço de abertura do pregão anterior	Numérica
GT	Volume de pesquisa do dia anterior do *Google Trends* através do nome da empresa	Numérica

Fonte: Elaboração própria.

Para realizar a modelagem, foi utilizada a linguagem de programação R com as seguintes bibliotecas: *class (Venables WN;2002), caret(Kuhn.; 2008), randomForest.(A. Liaw e M. Wiener; 2002).*

3.3 Análise dos dados

Os modelos são do grupo de algoritmos supervisionados de classificação e são utilizados no presente trabalho com o intuito de classificar o sentido do retorno para cada empresa, além do Ibovespa.. Diante disso, os modelos utilizados foram: Florestas Aleatórias, Regressão Logística, *Bagging*, *Boosting*, *KNN*, Máquinas de Vetores de Suporte (SVM) e o *XGBoost.*

3.3.1 Florestas aleatórias

As florestas aleatórias são um modelo de aprendizagem de máquina supervisionado que é composto por diversos modelos individuais de árvores de decisão, Basak *et al*. (2017). Assim, cada "árvore" da floresta é composta por um modo de decisão diferente, que classificou se o retorno foi positivo ou não. Para isso, cada modelo não deve ser correlacionado e os parâmetros dos modelos são a quantidade de galhos por árvore de decisão e o número de árvores que será composta na floresta.

3.3.2 Regressão Logística

A regressão Logística é um modelo linear generalizado que é representado da seguinte forma:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \qquad (2)$$

Em que $p(x)$ é uma probabilidade de o evento ser positivo, enquanto as demais variáveis x_i são as variáveis independentes (atributos), selecionadas para o estudo e a probabilidade $p(x)$ é que dirá a probabilidade de um valor ser da classe x.

Aplicando uma propriedade logarítmica, tem-se que:

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{p-1} x_{ip-1}}}{1+ e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{p-1} x_{ip-1}}} \qquad (3)$$

Assim, computa-se a probabilidade de o retorno ser positivo por meio do método de estimação de máxima verossimilhança, onde, o modelo é eficaz na predição de classes, Lee *et al.* (2018).

3.3.3 Bagging

Bagging é um método de aprendizagem de máquina, também conhecido como "agregado de bootstrap", cujo procedimento se reduz ao treinamento de árvores de classificação independentes, dada uma reamostragem com reposição, Nabipour *et al.* (2020). Após os classificadores serem treinados, o algoritmo irá combinar os demais classificadores, a fim de produzir o melhor classificador.

Neste estudo, será utilizado o seguinte pseudocódigo para a criação do modelo:

0. Seja x um vetor covariável;
1. Seja t um subconjunto de x reamostrado por bootstrap;
2. Considere C_t como classificadores configurados através de t;
3. Repita os passos 1 e 2 até t_n ;
4. Cada classificador será rankeado pela função a seguir:

$$C(x) = T^{-1} \sum_{t=1}^{T} C_t(x) \qquad (4)$$

em que x representa a base de dados original antes da reamostragem, T um subconjunto de x reamostrado por bootstrap e C uma função que irá classificar as amostras.

Assim, pode-se escolher o melhor classificador para o problema.

3.3.4 Boosting

A técnica de Boosting é utilizada para incrementar e aprimorar classificadores não tão robustos em sua classificação, Nabipour *et al. (2020)*. Para isso, atribui pesos e ordena-os de forma decrescente.

Seja x um vetor de entrada de dados e y o vetor de classificação correta de x, então uma distribuição pode ser assumida para os n números de x, onde pode-se predizer e

assume-se que os valores assumidos pelo preditor P são aqueles que se igualam a y; podemos chamar de E aqueles diferentes de y. Logo, pesos são atribuídos para os preditores, afim de encontrar aquele com o menor erro.

Já o *AdaBoost* tem como diferença que a distribuição D de x é normalizada.

3.3.5 Máquinas de Vetores de Suporte

É um modelo de aprendizagem de máquina supervisionado em que, conforme descrito por Moura *et al.* (2016), com duas classes de dados de entrada, é possível construir um hiperplano que as separe, para então, um classificador classificar novos dados de acordo com a posição relativa no hiperplano. Busca, assim, maximizar a distância entre as duas classes, caso o problema seja linear. Partindo do pressuposto de otimização, podemos escrever a equação do modelo desta forma:

$$\min_{\theta} C = \sum_{i=1}^{m}\left[y^{(i)} cost_1(\theta^T x_i) + (1 - y_i) cost_0(\theta^T x_i)\right] + \frac{1}{2}\sum_{i=1}^{n}\theta_j^2 \qquad (5)$$

Em que m é o número de elementos do vetor de treino, $cost_0$ e $cost_1$ são as funções de custo associadas à pertinência de uma amostra à classe ou não. y_i é a variável de saída para a amostra i, x_i é o vetor de variáveis de entrada para a amostra i, θ^Té o vetor de parâmetros transposto e o termo final é o termo de regularização para lidar com o problema de *overfitting.*

3.3.6 K- vizinhos próximos

Os *K*-vizinhos próximos, ou *KNN* do inglês, constituem um método de aprendizagem de máquina supervisionado utilizado para classificação. Para este trabalho, será assumido o vetor X_{ij} representando o vetor covariável com os dados de treino.

O modelo a ser efetuado para classificação depende da distância euclidiana (equação 8) entre as observações i e i_{+1} até i_n, e do classificador cujo critério de decisão será o da classe com maior probabilidade cuja equação (9) está detalhada a seguir:

$$C = \sqrt{\sum(x - x_1)^2} \qquad (6)$$

Em que x representa a variável que será classificada e x_i as demais observações do banco de dados.

$$\Pr(y = j \mid X = x_0) = \frac{1}{K}\sum_{i\epsilon N_0} l(Y_i = j) \qquad (7)$$

A probabilidade de um objeto y ser da classe j será dada através da relação com os K-vizinhos, Alkhatib *et. al*(2013).

Nesse caso, o modelo irá calcular a distância entre a nova observação e as K observações mais próximas, a fim de classificar a nova observação. A classe escolhida para a nova observação será a que possuir a maior frequência dentre os K vizinhos mais próximos.

3.3.7 XGBoost

O *XGBoost (eXtreme Gradient Boosting)* é um modelo da família das árvores de decisão, porém seu algoritmo funciona como uma técnica de agrupamento visto em Chen e Guestrin (2016), seleciona diversas árvores com pouca capacidade classificatória e então treina novos modelos com base nas árvores anteriores. O algoritmo é composto por diversos parâmetros, em que o modo como eles foram encontrados será discutida posteriormente, no tópico de avaliação de modelos.

Os parâmetros são:

max_depth = o máximo de profundidade das árvores, utilizado para controlar sobreajuste;

min_child_weight = o mínimo de somatório dos pesos, sendo utilizado para controlar o sobreajuste;

subsample = determina a fração de cada observação a ser selecionada randomicamente para criação de novas árvores, quanto menor o valor, mais conservador o algoritmo tende a ser, porém, leva para subajuste;

colsample_bytree = determina a fração de cada variável selecionada randomicamente para criação de novas árvores;

eta = determina a taxa de aprendizagem do modelo, quanto menor for o valor, mais robusto é o modelo, porém demanda elevado poder computacional, visto que com este parâmetro, irá demorar mais para encontrar o mínimo global para cada árvore.

3.4 Avaliação de modelos

Para validar o grau de performance de classificação, será necessário fazer a validação cruzada. Este procedimento foi feito no próprio treinamento de cada modelo citado anteriormente, através de *10-folds*. Ou seja, separou-se a amostra de treino em 10 subamostras e treinou-se o modelo em cada partição. Como cada modelo necessita da imputação de seu parâmetro, implementou-se uma *grid_search* a partir do pacote *caret*, com a finalidade de buscar o conjunto de parâmetros ótimos para cada modelo. Dessa maneira, buscou-se minimizar o erro para a amostra de validação e, em seguida, foram selecionados os modelos finais para o ativo. Posteriormente, foi preciso prever o modelo com a amostra de teste, a fim de avaliar a capacidade de generalização de cada modelo.

Diante disso, as métricas de avaliação final dos modelos foram:

- Acurácia: Mede o quanto o modelo conseguiu prever corretamente;
- Curva *ROC:* é utilizada para visualizar classificadores de acordo com suas performances.

Para demonstrar o desempenho dos modelos, será utilizada a Matriz de Confusão, que mede a frequência de classificação para cada modelo.

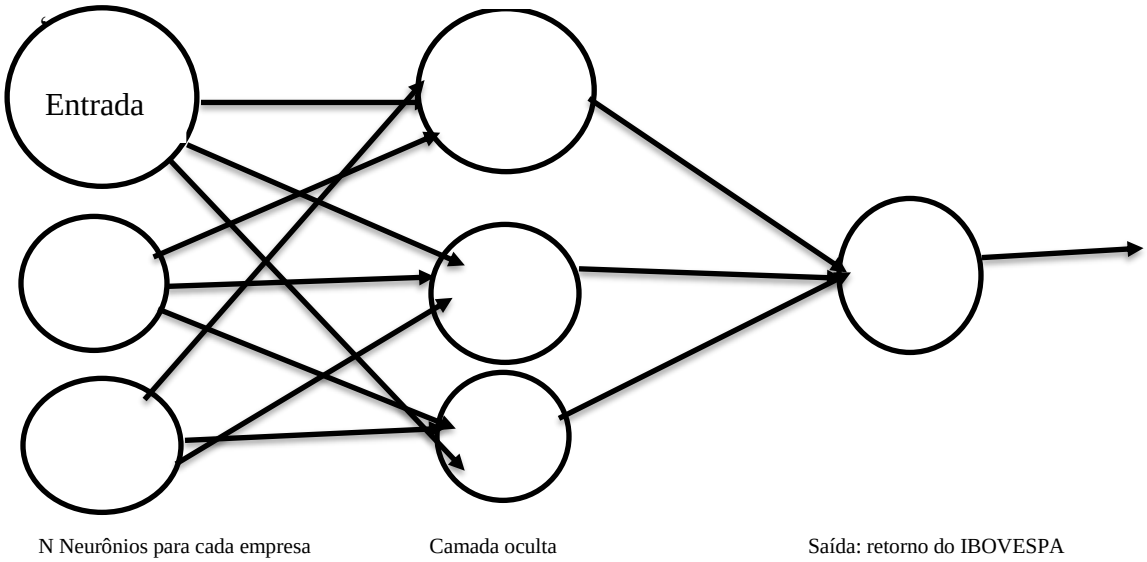
3. 5. Redes neurais artificiais

As Redes Neurais Artificiais (RNAs) são estruturas baseadas nos neurônios do cérebro humano. As RNAs são divididas em três grandes grupos: entrada, camada oculta e saída, conectados por nódulos. A camada de entrada são as informações referentes às características ou variáveis preditoras (e.g., *Google Trends*, logaritmo do volume negociado, preço de abertura do índice e das empresas, preço de fechamento do índice e das empresas); a camada oculta é composta por neurônios onde são feitas as demais combinações dos diferentes dados de entrada; e a camada de saída é composta pela variável a ser prevista, ou seja, a direção do Ibovespa no dia seguinte.

Nos próximos parágrafos serão demonstrados a estrutura da rede neural.

Uma rede neural, matematicamente, é um somatório de dados de entrada ponderados por pesos e caso as condições da função de ativação forem ativadas, terá a saída necessária, a Figura 1 exemplifica de forma esquemática uma rede neural.

Figura 1: Estrutura da rede neural



Fonte: elaboração própria.

A rede neural descrita na imagem possui apenas uma camada oculta, que, apesar de ser bastante eficiente para classificar na literatura, assume uma separabilidade linear entre as variáveis no hiperplano. Por conta disto, foi utilizada nesta monografia a *multilayer perceptron*, onde diferente *da singlelayer* perceptron, os seus pressupostos são descritos conforme observado em Haykin (2008):

- O modelo de cada neurônio inclui uma função de ativação não linear que é diferenciável;
- A rede contém mais de uma camada que são ocultas tanto para os dados de entrada, quanto para os dados de saída, e;
- A rede exibe um elevado grau de conectividade, a amplitude de cada neurônio é determinada por pesos sinápticos da rede.

Visto que o problema a ser resolvido é de classificação, a função de ativação da RNA é a função sigmoidal ou logística. Os parâmetros que foram selecionados para a busca do *grid* foram decaimento do peso dos neurônios, e, quantidade de neurônios na camada oculta.

3.6 Comparação de estratégias de investimento

Ao final do estudo foi efetuado o *backtest* entre os modelos que tentaram prever o Ibovespa, mensurando o retorno logarítmico de uma estratégia teórica, onde: caso no próximo dia útil o índice fosse ter retorno positivo ao final do dia, o investidor entraria comprado no índice; caso o índice fosse ter retorno negativo, ou neutro, o investidor venderia antes do pregão. Para validar o resultado, comparou-se o retorno da estratégia de investimento descrita acima com a do *Buy and Hold*, que é uma estratégia conservadora de investimentos, onde compra-se e segura a ação, sem ocorrência de venda.

4 RESULTADOS

4.1 Estatísticas descritivas

A Tabela 1 demonstra as estatísticas descritivas por ativo, para cada variável analisada. Observando os números, pode-se verificar que a variável *value* (máximo de valores pesquisados) foi de 100 pontos, e o mínimo, para a maioria dos ativos, foi de 0 pontos. Porém, a Ambev, a *holding* Itaúsa e a mineradora Vale não apresentaram pontuação zero como valor mínimo ao longo dos anos. De forma a se analisar o risco das empresas, dado pelo desvio padrão do retorno, pode-se verificar que a empresa mais arriscada para se investir durante o período foi a Petrobrás, onde apresentou um risco aproximado de 3,1%, o que se relaciona também, pela crise política marcada pelo período. Já a empresa com o menor risco, foi a Ambev, com risco aproximado de 1,4%.

Apesar do retorno médio diário dos ativos serem zero, devido à natureza estacionária da variável, o período analisado de 2012 até 2018 foi marcado por bastante incerteza no Brasil, conforme estudado por Barboza e Zilberman (2018): entre 2014 e 2017, ocorreram duas quebras de máximas em relação ao índice de Incerteza no Brasil, ocorrendo uma crise financeira e política, elevando o grau de incerteza para os investimentos no país. As demais informações sobre as variáveis são encontradas na Tabela 1.

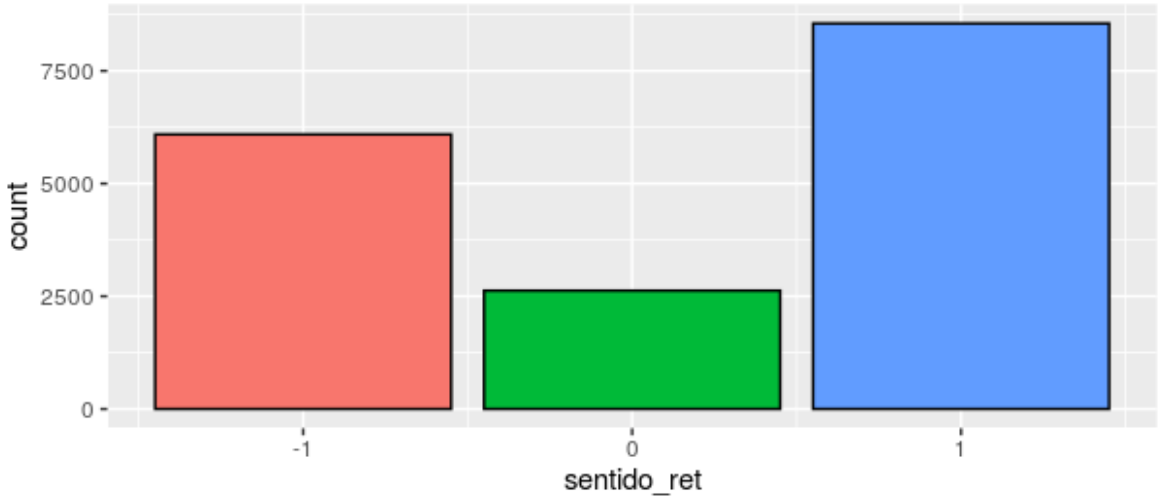
Tabela 1: Estatísticas descritivas dos nove ativos e do Ibovespa referentes ao período de 2012 a 2018

Ativo	Ret medio	Ret mediano	Ret sd	Ret min	Ret max	Vlm medio	Vlm mediano	Vlm sd	Vlm min	Vlm max	Value med	Value mediano	Value sd	Value min	Value max
ABEV3	0,000	0,000	0,014	-0,058	0,107	11,802	12,248	1,164	8,426	14,240	62,096	62,000	13,461	27,000	100,000
B3SA3	0,001	0,000	0,021	-0,088	0,096	12,122	12,111	0,505	9,581	14,045	4,831	0,000	14,985	0,000	100,000
BBDC4	0,000	0,000	0,020	-0,141	0,122	12,704	12,697	0,409	10,261	14,238	35,614	34,000	22,342	0,000	100,000
BBAS3	0,000	0,000	0,027	-0,238	0,134	12,364	12,340	0,507	10,088	14,343	43,185	39,000	20,494	0,000	100,000
ITSA4	0,000	0,000	0,019	-0,101	0,098	12,051	12,040	0,408	9,981	13,609	69,505	70,000	11,440	30,000	100,000
ITUB4	0,000	0,001	0,019	-0,128	0,104	13,068	13,051	0,419	10,836	14,985	32,412	33,000	21,167	0,000	100,000
PETR3	0,000	0,000	0,031	-0,162	0,150	12,184	12,164	0,571	9,536	14,440	36,703	34,000	20,313	0,000	100,000
PETR4	0,000	0,000	0,031	-0,171	0,151	13,514	13,466	0,512	11,413	15,442	40,564	39,000	18,667	0,000	100,000
VALE3	0,000	0,000	0,027	-0,157	0,138	12,362	12,208	0,785	9,753	14,905	73,812	78,000	13,456	38,000	100,000
IBOV	0,000	0,000	0,014	-0,092	0,064	15,937	15,925	0,318	14,178	17,330	31,760	32,000	19,342	0,000	100,000

Fonte: elaboração própria.

Em termos visuais, o Gráfico 2 demonstra a quantidade de vezes em que o retorno foi neutro, negativo, ou, positivo. De forma suscinta, o gráfico a seguir representa de forma a contar as informações dos retornos das empresas.

Gráfico 1: Contabilização do sentido do retorno diário das empresas, referente ao período de 2012 a 2018?



Fonte: elaboração própria.

Observando o gráfico, é possível visualizar que as empresas, no geral, tiveram mais retorno positivo que retorno negativo ao longo dos anos. Para facilitar a modelagem, o fator -1, foi enquadrado no fator 0, assim, o trabalho ficou com duas classes para serem modeladas.

Já em relação à quantidade de buscas pelo *Google Trends*, o Gráfico 3 demonstra a variável por ativo, além do Ibovespa, em relação ao tempo. Assim, foi possível visualizar que, cerca de 7500 dias as empresas tiveram retorno positivo, e, aproximadamente 7500 com o somatório de dias com retornos neutros e negativos. Assim, tem-se uma categorização bastante ampla da variável alvo.

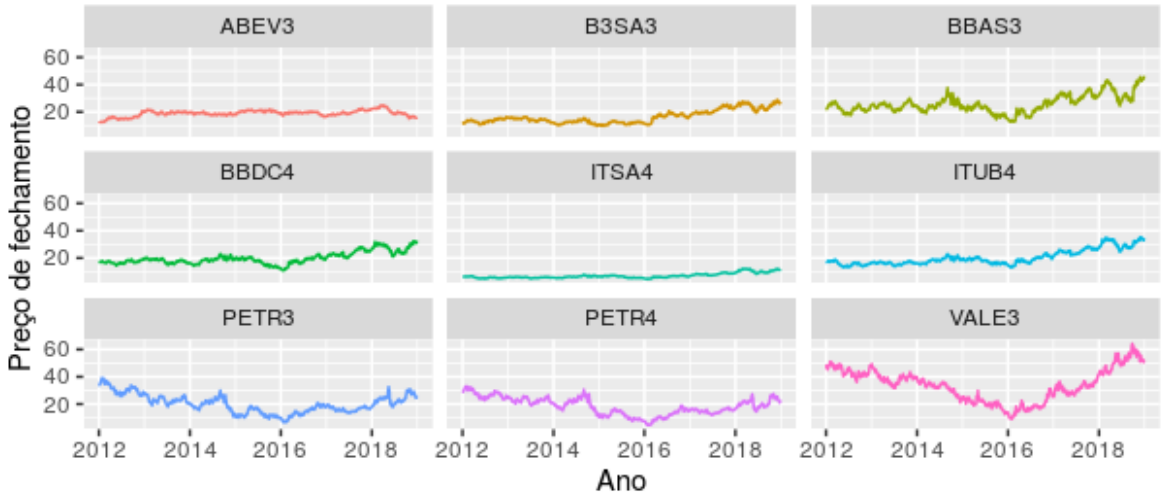
A seguir, o Gráfico 3 demonstra a frequência de busca pelas empresas e pelo Ibovespa durante o período analisado:

Gráfico 2: Frequência do índice de busca Google Trends durante o período estudado, entre 2012 e 2018, para os nove ativos e o Ibovespa

Fonte: elaboração própria.

No que diz respeito à variável *Google Trends*, há bastante volatilidade para os ativos, devido ao elevado desvio padrão, valores máximos e mínimo, menos com a empresa B3SA3, por ter aberto capital em 2017. A ABEV3, ITSA4 e VALE3, por serem empresas de bastante relevância nacional, e, internacional, acabaram não apresentando valor mínimo zero na busca para o Google Trends. Visualizando o fechamento de cada ativo e o do IBOVESPA:

Gráfico 3: Preço dos fechamentos dos nove ativos durante o período de 2012 e 2018



Fonte: elaboração própria.

No Gráfico 4, foi possível observa a trajetória dos preços das empresas ao longo do período estudado. Aproximadamente as nove empresas tiveram seus preços das ações reduzidas entre 2014 e 2016, período onde houve a crise política brasileira. Poucas empresas se mantiveram sólidas, salientando a ITSA4, que, por ser uma *holding* e deter participação em outras companhias, manteve se manteve relativamente constante durante a crise. A seguir, o Gráfico 5 representa o fechamento do índice de mercado da bolsa de valores brasileira:

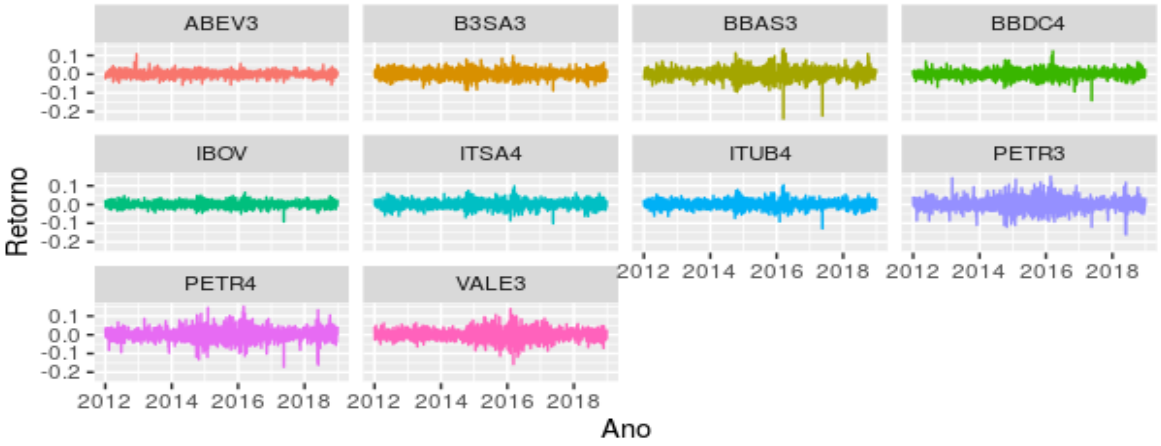
Gráfico 4: Fechamento do índice Ibovespa entre o período estudado de 2012 e 2018.

Fonte: elaboração própria.

Em ambos os gráficos, é possível verificar dois períodos distintos o período de pré-recessão financeira entre 2012 e início de 2014, e, o período pós-recessão, entre meados de 2014 até o final de 2018 ,onde tanto o IBOV, quanto a maioria das empresas, tiveram a elevação dos seus preços.

O Gráfico 6 representa o retorno dos 9 ativos e do Ibovespa ao longo do período entre 2012 e 2018:

Gráfico 5: Retorno diário dos nove ativos e Ibovespa referentes ao período entre 2012 e 2018.



Fonte: elaboração própria.

Já os retornos, os mesmos se mantiveram estacionários durante o período, isto é, ao longo do tempo, suas variâncias e médias, foram zero e constante, respectivamente, conforme a visualização do Gráfico 6.

A resumir, nesta seção foram as descrições das variáveis, salientando para o risco elevado da Petrobrás durante o período, retornos constantes, indicando estacionariedade e por fim, foi possível visualizar que nos sete anos de análise, as ações com maior volume negociados, oscilaram num período pré-crise, entre 2012 e 2016, e em um período pós-crise, de 2016 e 2018.

4. 2. Modelagem preditiva

Foram testados os modelos para todas as empresas, porém, pelos resultados não terem sido satisfatórios - pois a taxa de acurácia média dos modelos testados foi de 50,04%, o que em termos de predição, faz com que a probabilidade da ação ou subir no dia seguinte, seja de 50% – não foi optado por utilizar a predição das empresas para a predição do IBOVESPA. Os resultados podem ser encontrados na tabela 2 a seguir:

Tabela 2: Resultados da modelagem por empresa

Modelos								
Empresas	**ADABOOST**		**BAGGING**		**KNN**		**REGRESSÃO LOGÍSTICA**	
	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança
ABEV3	0,5061	0,4417 - 0,5704	0,498	0,4337 - 0,5623	0,4735	0,4096 - 0,538	0,5265	0,462 - 0,5904
B3SA3	0,4939	0,4296 - 0,5583	0,5102	0,4458 - 0,5744	0,5714	0,5069 - 0,6342	0,4857	0,4216 - 0,5502
BBAS3	0,449	0,3856 - 0,5136	0,4853	0,4216 - 0,5502	0,5184	0,4539 - 0,5824	0,5061	0,4417 - 0,5704
BBDC4	0,449	0,3856 - 0,5136	0,5306	0,446 - 0,5944	0,5102	0,4458 - 0,5744	0,4939	0,4296 - 0,5583
ITSA4	0,5388	0, 4742 - 0,6024	0,5306	0,446 - 0,5944	0,5143	0,4498 - 0,5784	0,502	0,4377 - 5663
ITUB4	0,449	0,3856 - 0,5136	0,5102	0,4458 - 0,5744	0,5143	0,4498 - 0,5784	0,5143	0,4498 - 0,5784
PETR3	0,4571	0,3936 - 0,5218	0,4857	0,4216 - 0,5502	0,4735	0,4096 - 0,538	0,5184	0,4539 - 0,5824
PETR4	0,4653	0,4016 - 0,5299	0,4898	0,4256 - 0,5542	0,4857	0,4216 - 0,5502	0,4816	0,4176 - 0,5461
VALE3	0,4653	0, 4016 - 0,5299	0,4776	0,4136 - 0,5421	0,4694	0,4056 - 0,534	0,502	0,4377 - 0,5663

Modelos								
Empresas	**Random Forest**		**SVM LINEAR**		**SVM RADIAL**		**XGBOOST**	
Métricas	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança	Acurácia	Intervalo de Confiança
ABEV3	0,5429	0,4782 - 0,6064	0,5306	0,466 - 0,5944	0,5347	0,4701 - 0,5984	0,5551	0,4905 - 0,6184
B3SA3	0,5388	0,4742 - 0,6024	0,502	0,4377 - 0,5663	0,502	0,4377 - 0,5663	0,502	0,4377 - 0,5663
BBAS3	0,4776	0,4136 - 0,5421	0,4653	0,4016 - 0,5299	0,4735	0,4096 - 0,538	0,5143	0,4498 - 0,5784
BBDC4	0,502	0,4377 - 0,5663	0,4816	0,4176 - 0,5461	0,4816	0,4176 - 0,5461	0,4735	0,4096 - 0,538
ITSA4	0,5061	0,4417 - 0,5704	0,498	0,4337 - 0,5623	0,4776	0,4136 - 0,5421	0,5347	0,4701 - 0,5984
ITUB4	0,502	0,4377 - 0,5663	0,498	0,4337 - 0,5623	0,5102	0,4458 - 0,5744	0,4816	0,4176 - 0,5461
PETR3	0,4816	0,4176 - 0,5461	0,4857	0,4216 - 0,5502	0,4857	0,4216 - 0,5502	0,4857	0,4216 - 0,5502
PETR4	0,4653	0,4016 - 0,5299	0,502	0,4377 - 0,5663	0,5102	0,4458 - 0,5744	0,4939	0,4296 - 0,5583
VALE3	0,4735	0,4096 - 0,538	0,5061	0,4417 - 0,5704	0,5061	0,4417 - 0,5704	0,4816	0,4176 - 0,5461

Fonte: Elaboração própria.

De forma a resumir a Tabela 2 demonstra a modelagem por ativo durante o período analisado, visto isso, a percebe-se que de fato os modos não tiveram alta capacidade preditiva, porém, tal fato pode ser justificado pelo período de instabilidade da bolsa de valores brasileira, e também pelo fato das ações se comportarem como um *Random Walk*, ou seja, se comporta de forma aleatória..

Para a predição do Ibovespa, foram testados os modelos *Xgboost* e Rede Neural Artificial com uma única camada oculta. O número de neurônios encontrado na *Grid* foi de 9, o decaimento dos pesos foi de 0,91, e a função de ativação foi sigmoidal. Os resultados estão demonstrados na Tabela 3.

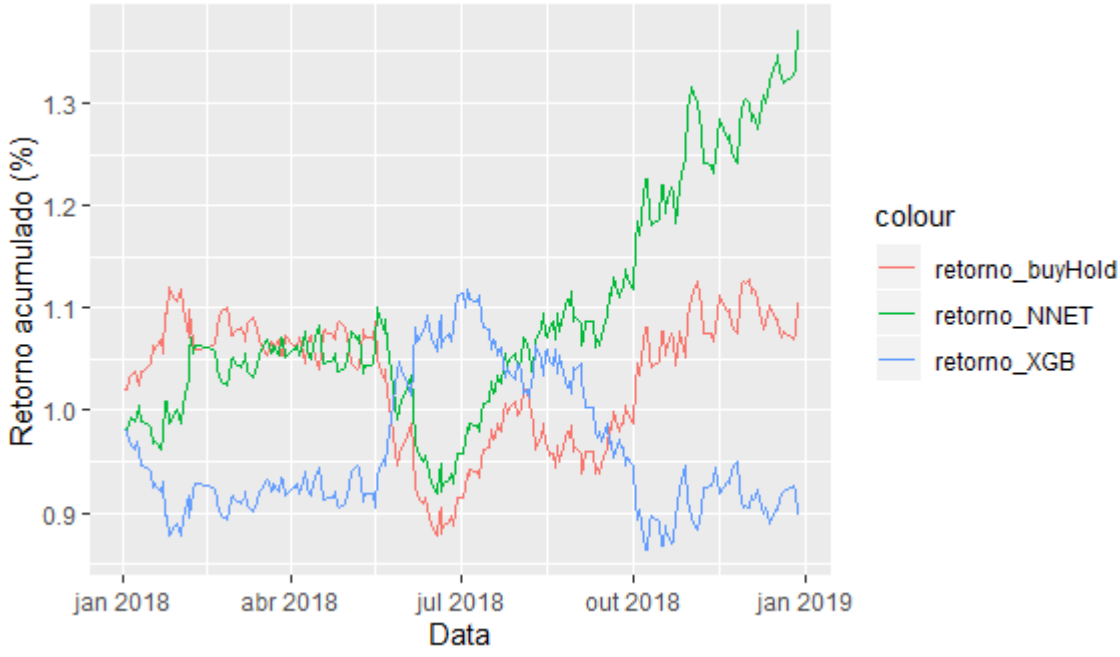
Tabela 3: Comparação entre modelos de previsão do Ibovespa: Rede Neural e *XgBoost*

Modelo	Acurácia	Intervalo de Confiança
Rede Neural	0,5306	0,466 - 0,5944
XGBOOST	0,4939	0,4296 - 0,5583

Fonte: elaboração própria.

Apesar dos resultados não serem igualmente satisfatórios para a predição do sentido do Ibovespa, tendo uma probabilidade igual ao limite do ponto de corte (50%), pois, conforme dito anteriormente, a probabilidade de o índice subir ou descer no dia seguinte, é meia-a-meia. O Gráfico 7 apresenta o *Backtest* de comparação da estratégia, realização de lucro ou prejuízo, durante o período de 2018.

Gráfico 6: *Backtest* entre estratégias de investimento



Fonte: elaboração própria.

Visualizando o Gráfico 7, percebe-se que, no período em que houve o início da corrida eleitoral brasileira, em julho, a estratégia proposta é capaz de obter retornos maiores que o método tradicional de investimentos. Então, após durante o período eleitoral até o final do ano, os retornos da estratégia pela rede neural foram superiores ao do Buy and Hold, pois a rede neural é capaz de captar informações, e, encontrar falhas no mercado, em discordância com o que é dito na Hipótese de Mercados Eficientes, é possível observar que nem sempre toda a informação disponível é precificada pelos mercados. Tal resultado pode ser correlacionado com a instabilidade do período, dado que o ano de 2018 foi um ano atípico no mercado, por conta de incerteza política, e, perspectivas de retomada da economia.

5 CONCLUSÃO

A modelagem preditiva supervisionada foi utilizada para a predição do sentido do retorno das empresas, e, a aprendizagem profunda, por meio das redes neurais, para prever o sentido do Ibovespa. O presente estudo teve como objetivo utilizar as informações do *Google Trends* e os algoritmos de aprendizado de máquina para prever a direção do retorno das empresas com maior volume negociado na B3 entre 2012 e 2018 e então prever o sentido do Ibovespa. Além do exposto, o estudo buscou trazer para a pesquisa financeira brasileira a inclusão da variável *Google Trends*, que por meio de compilação de estudos, tem se mostrado relevante segundo pesquisas nacionais e internacionais.

Os resultados observados não foram satisfatórios, de modo que o intervalo de confiança encontrado foi em torno de 45%-50%, o que não torna uma predição acurada, em termos de tomada de decisão de investimento. Porém, ao olharmos para o retorno acumulado do período de testes, a estratégia proposta com 1,37% de retorno, teve um retorno maior em relação à estratégia conservadora de investimentos, com 1,10%, nesse sentido, a modelagem proposta pode auxiliar fundos de investimentos ativos a manterem uma performance superior ao do Ibovespa. Tal resultado pode ser justificado pelo fato de a rede neural perceber padrões do mercado, onde um ser humano não conseguiria captar com a mesma velocidade e precisão do modelo, mesmo que a acurácia não seja maior que o ponto de corte de 50%.

Para estudos futuros, fica proposto o ajustamento das empresas para cada distribuição estatística específica, e o uso de outros modelos de aprendizagem de máquina e a classificação em três classes, não apenas binário. Ademais, é sugerida também, treinar durante os anos de 2017 e 2018, e testar os modelos em 2019, em um período pós-crise econômica e política brasileira.

REFERÊNCIAS

ALKHATIB, K.; NAJADAT, H.; HMEIDI, I.; SHATNAWI, K. A. M. Stock Price prediction using K-Nearest Neighbor (kNN) Algorithm**. International Journal of Business, Humanities and Technology**. v. 3 n. 3. 2013.

BARBOZA, M. de. R; ZILBERMAN, E.; Os Efeitos da Incerteza sobre a Atividade Econômica no Brasil. **Revista Brasileira de Economia.** v. 72 n.2. Brasil. Junho. 2018.

BASAK, S.; KAR, S.; SAHA, S.; KHAIDEM, L.; DEY, R. S.; Predicting the direction of stock market prices using tree-based classifiers. **North American Journal of Economics and Finance**. v. 06 n. 13, 2018.

CARIDE, M. I.; BARIVIERA, A. F.; LANZARINI L. Stock Returns Forecast: An Examination By Means of Artificial Neural Networks. **Studies in Systems Decision and Control.** v. 125. p. 399-410. 2018.

CARNEIRO, A. H; MYLONAKIS, E. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. **Clinical Infectious Diseases:** an official publication of the infectious diseases Society of America**,** 49 (10). 2009. p. 1555-1564.

CHEN, T.; GUESTRIN, C.; XGBoost: A Scalable Tree Boosting System. **Expert Systems With Applications**. v. 09. n. 005. 2018.

CHEN, Y.; HAO, Y. A feature wighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. **Expert Systems with Applications**. v. 02 p. 044. 2017

CHOI. H.; VARIAN. H., Predicting the present with Google Trends. **Economic Record**. v.88, n. s1, 2012. p.2-9.

ETTREDGE, M.; GERDES, J.; KARUGA, G.; Using Web-Based search Data to Predict Macroeconomic Statistics. **Communications of the ACM**. v. 48, n. 11, 2005. p. 87-92.

FAMA, E. F., Efficient Capital Markets: A Review of Theory and Empirical Work , **The Journal of Finance**. v. 25, n. 2, 1970. p. 383-417.

FAMA, F. E.; FRENCH, R. K. Size and Book-to-Market Factors in Earnings and Returns. **The journal of Finance**. v. 50, n. 1, 1995. p. 131-155.

FONDEUR, Y.; KARAMÉ., F. Can Google data Help predict French youth unemployment? **Economic Modelling.** v. 30, 2013. p. 117-125.

FUKUSHIMA, K.; Neocognitron: A Self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. **Biological Cybernetics**. v. 36, n. 4. pp 193-202. 1980.

GRIFFITH, A. K.; A comparison and evaluation of three machine learning procedures as applied to the game of checkers. **Artificial Intelligence**, *5(2), 137–148.* 1974.

HAYKIN, S.; Neural Networks and Learning Machines. 3ª edição, Estados Unidos. Pearson, 2008.

HILL, T.; MARQUEZ, L.; O'CONNOR, M.; REMUS, W. Artificial neural network models for forecasting and decision making. International Journal of Forecasting. v. 10 n. 1 p. 5-15. 1994.

KIM, N. Luˇcivjansk´a, K., MOLNÁR, P., VILLA, R. Google searches and stock market activity: evidence from Norway. **Finance Research Lab.** . v. 28, n. dd, 2018.p. 208-220.

KOHLI, S. P. P.; ZARGAR, S.; ARORA, S.; GUPTA, P.; Stock prediction Using Machine Learning Algorithms. **Applications of Artificial Intelligence Techniques in Engineering**. v. 698, 2018. p. 405-414.

KRISTOUFEK. L. Can Google Trends search queries contribute to risk diversification? **Scientific Reports**. v. 3, 2013.

KRISTOUFEK, L. MOAT, S. H., PREIS, T. Estimating suicide occurrence statistics using Google Trends. **EPJ Data Science**. v. 5, n. 32, 2016.

LEE, T. K; CHO, J. H.; KWON, D. S.; SOHN, S. Y. Global stock market investment strategies based on financial network indicators using machine learning techniques. **Expert Systems with Applications**. v. 09 p. 005. 2018.

LINTNER J., The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. **The Review of Economics and Statistics**. v. 47, n. 1, 1965. p. 13-37.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin Of Mathematical Biophysics**, New York, p. 115-133. dez. 1943.

MACHADO, J. E., OLIVEIRA, R., PEREIRA, M. C. A. Proposal and Implementation of machine learning and deep learning models for stock markets using web data. 2015.

MARQUES, S., AHFELDT, R., CRUZ, W. A. J., SILVA, V. W. Análise de anomalia da hipótese dos mercados eficientes à luz das finanças comportamentais. **Revista da Faculdade de Administração e Economia**. v. 6, n. 2, p. 33-50. 2015.

MARKOWITZ, H., Portfolio Selection, **The Journal of Finance**. v. 7. n. 1. 1952. 77-91

MARRETTI, R.; OMAR, N.. RANDOM FOREST APLICADO AO MERCADO BRASILEIRO DE AÇÕES. **CONTECSI USP - International Conference on Information Systems and Technology Management - ISSN 2448-1041**, Brasil, abr. 2019.

MOURA de, M. S. F. I., MORAES, R. M., MACHADO dos, S. L. Avaliação para Trajetórias de Incisões Cirúrgicas com *SVM*. **Revista de Informática Aplicada.** 2016.

NABIPOUR, M.; NAYYERI, P.; JABANI, H.; SHAMSHIRBAND, S.; MOSAVI, A. Deep Learning for Stock Market Prediction. **Preprints**. v. 1. 2020.

NODA F. R., MARTELANC R., KAYO K. E. O fator de Risco Lucro/Preço em Modelos de Precificação de Ativos Financeiros. **Revista de Contabilidade e Finanças.** v. 27(70). 2014. p. 67-79.

PAGOLU, S. V., REDDY. N. K., PANDA, G., Majhi, B. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016. Paralakhemundi. pp. 1345-1350.

PAIVA, D. F., Modelos de precificação de ativos Financeiros de fator único: Um teste empírico dos modelos CAPM e D-CAPM. **Caderno de pesquisas em administração**. v. 12 2005. pp. 49-65.

PATEL, J., SHAH, S., THAKKAR, P., KOTECHA, K. Predicting stock market Index Using fusion of machine learning techniques. **Expert Systems with Applications. v. 42. n. 4.** 2014. pp. 2162-2172.

PEDREGOSA, F., VAROQUAUX, G. GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., DUCHESNAY, E. Scikit-learn: Machine Learning in Python. J**ournal of Machine Learning Research.** v. 12. 2011. pp. 2825-2830.

ROSENBLATT, M. The Perceptron: A probabilistic model for information storage and organization in the Brain. **Psychological review**, v.65, n.6, p. 386-408. 1958.

SAMUEL. A. L., Some Studies in Machine Learning Using the Game of Cheekers. **IBM J. RES. DEVELOP.** 3.3. 1959. 210-229.

SANTOS DOS. J., FAMÁ A., MUSSA A. A adição do Fator de risco momento ao modelo de precificação de ativos dos três fatores de Fama & French aplicado ao mercado acionário brasileiro. **Revista de Gestão**. v. 19. n. 3 2011. pp. 453-471.

SHARPE, F. W., Mutual Fund Pefomance. **The Journal of Business**. 39 (1)1966. pp. 341-360.

SHARPE, F. W. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. **The Journal of Finance**. v. 19. n. 3. 1964. pp. 425-442.

SILVA, V. C; BATISTA, N. T. A.; SALES, L. H; da PENHA, S. R.; Aplicação do modelo Monte Carlo Na avaliação da Empresa Ambev com Custo de Capital impreciso. **Revista ENIAC**. v. 8. n. 1. 2019. pp 153-175.

SWALLOW-CARRIÈRE, Y.; LABBÉ. F.; Nowcasting with Google Trends in an Emerging Market. **Journal of Forecasting. v**. 32. 2011. pp. 289-298.
TAYLOR, W. K.; Machine learning and recognition of faces. **Electronics Letters**, 3(9), 436. 1967.

TSAI, F. C., WANG, P. S. Stock Price Forecasting by Hybrid Machine Learning Techniques. **Proceedings of the International MultiConference of Engineers and Computer Scientists**. v. 1. 2009. Hong Kong.

VARGAS, M. R., de LIMA, B. S. L. P., EVSUKOFF, A. G. *Deep learning for stock market prediction from financial news articles.* **International Conference on Computational Intelligence and Virtual Environments for measurement Systems and Applications.** Annecy. 2017. pp. 60-65.

XIONG, R. NICHOLS, P. E., SHEN, Y. Deep Learning Stock Volatility with Google Domestic Trends. **Google.** 20