

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

MARCELO RIVERA MARIN

MODELO DE SELEÇÃO DE AÇÕES A PARTIR DE REGISTROS CONTÁBEIS E DE
INDICADORES DE MERCADO, UTILIZANDO TÉCNICAS DE ANÁLISE DE
COMPONENTES PRINCIPAIS, *SUPPORT VECTOR MACHINE* E REDES NEURAIS
NO MERCADO BRASILEIRO

SÃO PAULO

2020

MARCELO RIVERA MARIN

MODELO DE SELEÇÃO DE AÇÕES A PARTIR DE REGISTROS CONTÁBEIS E DE
INDICADORES DE MERCADO, UTILIZANDO TÉCNICAS DE ANÁLISE DE
COMPONENTES PRINCIPAIS, *SUPPORT VECTOR MACHINE* E REDES NEURAIS
NO MERCADO BRASILEIRO

Dissertação apresentada à Escola de
Economia de São Paulo da Fundação
Getulio Vargas como requisito para
obtenção do título de Mestre Profissional
em Economia.

Campo de conhecimento: investimentos.

Orientador: Prof. Dr. Ricardo R. Rochman

SÃO PAULO

2020

Marin, Marcelo Rivera.

Modelo de seleção de ações a partir de registros contábeis e de indicadores de mercado, utilizando técnicas de análise de componentes principais, *support vector machine* e redes neurais no mercado brasileiro / Marcelo Rivera Marin. - 2020.

62 f.

Orientador: Ricardo Ratner Rochman.

Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Ações (Finanças). 2. Mercado financeiro - Brasil. 3. Aprendizado do computador. 4. Redes neurais (Computação). 5. Análise de componentes principais. I. Rochman, Ricardo Ratner. II. Dissertação (mestrado profissional MPFE) – Escola de Economia de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 336.763.2(81)

MARCELO RIVERA MARIN

MODELO DE SELEÇÃO DE AÇÕES A PARTIR DE REGISTROS CONTÁBEIS E DE
INDICADORES DE MERCADO, UTILIZANDO TÉCNICAS DE ANÁLISE DE
COMPONENTES PRINCIPAIS, *SUPPORT VECTOR MACHINE* E REDES NEURAIS
NO MERCADO BRASILEIRO

Dissertação apresentada à Escola de
Economia de São Paulo da Fundação
Getulio Vargas como requisito para
obtenção do título de Mestre Profissional
em Economia.

Campo de conhecimento: investimentos.

Orientador: Prof. Dr. Ricardo R. Rochman

Data de aprovação: 26/05/2020

Banca examinadora:

Ricardo Ratner Rochman

(Orientador)

FGV – EESP

Elia Matsumoto

FGV – EESP

Gustavo Mirapalheta

FGV – EAESP

DEDICATÓRIA

Dedico esta dissertação a minha esposa Fernanda e as minhas filhas Gabriela e Rebeca. À Fernanda, seu sacrifício em minhas ausências, seu apoio incondicional, suas palavras de carinho e motivação foram essenciais nessa jornada. Às minhas filhas, pelas brincadeiras e motivações natas, sem contar pelas pressões e trapalhadas que fizeram deste um dos meus maiores desafios.

AGRADECIMENTOS

Ao Prof. Dr. Ricardo Rochman, meu orientador, por me ajudar a construir e desenvolver este trabalho. E por me orientar a tomar as decisões corretas nas encruzilhadas deste curso.

A SulAmérica Investimentos por apoiar os meus estudos. Agradeço ao Juan Morales por me ajudar com discussões construtivas e, principalmente, ao Philipe Biolchini, pelo seu incentivo e apoio nesta empreitada.

Por fim, e não menos importante, aos meus pais por todo o esforço feito para que eu conseguisse realizar os meus sonhos e, por estarem sempre ao meu lado, incondicionalmente.

RESUMO

O objetivo deste trabalho é propor um modelo de seleção de ações que superem o índice IBrX 100, com base em seus registros contábeis e indicadores de mercado, a partir da utilização das técnicas de *support vector machine* e redes neurais - identificadas ao longo deste estudo como classificadores.

Os classificadores são treinados em parte da amostra e seguem para a verificação de suas precisões a partir dos conceitos extraídos da matriz confusão. Em seguida, é analisado o desempenho da carteira formada pelos ativos selecionados do classificador com maior acurácia em relação ao índice IBrX 100.

O classificador *support vector machine* com kernel linear apresentou resultados superiores dentre os demais na identificação das ações vencedoras, com acurácia de 54,8%. Ao analisarmos as características da classificação desse algoritmo, nota-se que foram selecionadas as ações que apresentaram aumento de rentabilidade, redução de endividamento e maiores retornos nos períodos de 2 meses, 1 mês e 5 dias.

A carteira formada pela seleção de ações a partir do classificador vencedor desempenhou 17,8% melhor que o IBrX 100, com *information ratio* médio de 0,30 em 2018 e 1,48 em 2019. Como possível crítica deste resultado, a amostra de teste ficou concentrada em um período de tendência de alta do índice de referência.

Palavras-chave: seleção de melhores ações, classificadores, *support vector machine*, redes neurais, análise de componentes principais.

ABSTRACT

The goal of this work is to propose a stock selection model that overcomes the IBrX 100 index, based on its accounting records and market indicators, based on the use of support vector machine techniques and neural networks – identified throughout this study as classifiers.

The classifiers are trained in part of the sample and proceed to their precision verification based on the concepts extracted from the confusion matrix. Then, the performance of the portfolio formed by the selected assets of the classifier with greater accuracy in relation to the IBrX 100 index is analyzed.

The support vector machine classifier with linear kernel showed superior results among the others in the identification of winning stocks, with an accuracy of 54,8%. When analyzing the characteristics of the classification of this algorithm, it is noted that the stocks selected were the ones that showed increased profitability, debt reduction and higher historical returns.

The portfolio formed by the stocks selected by the winning classifier performed 17.8% better than the IBrX 100, with an average information ratio of 0.30 in 2018 and 1.48 in 2019. As a possible criticism of this result, the test sample was concentrated in a period of upward trend in the reference index.

Key words: stocks selection, classifiers, support vector machine, neural networks, principal component analysis.

LISTA DE FIGURAS

Figura 1: Elaboração da base de dados.....	18
Figura 2: Janela de estudo para cada ativo.....	21
Figura 3: Kaiser-Meyer-Olkin (KMO)	28
Figura 4: Exemplo de separador linear do hiperplano.....	29
Figura 5: Representação clássica de um modelo de neurônio artificial.....	31
Figura 6: Matriz confusão	36
Figura 7: Matriz confusão do SVM linear	45
Figura 8: Comparação das variáveis de rentabilidade	46
Figura 9: Comparação das variáveis de mensuração de valor.....	47
Figura 10: Comparação das variáveis de liquidez e endividamento.....	48
Figura 11: Comparação das variáveis de desempenho de mercado nas diferentes amostras.....	48
Figura 12: Desempenho contra o índice de referência nas previsões da amostra de teste	49
Figura 13: Desempenho acumulado da Carteira 1 e do índice de IBrX 100	51
Figura 14: Desempenho trimestral da Carteira 1 e do índice IBrX 100	53
Figura 15: Evolução da volatilidade da Carteira 1 e do IBrX 100	54
Figura 16: Evolução do risco ativo e do information ratio da Carteira 1	55

LISTA DE TABELAS

Tabela 1: Variáveis de rentabilidade.....	22
Tabela 2: Variáveis binárias.....	23
Tabela 3: Variáveis de liquidez e endividamento.....	23
Tabela 4: Variáveis de mensuração de valor.....	24
Tabela 5: Variáveis de desempenho de mercado.....	25
Tabela 6: Quantidade de neurônios nas camadas.....	33
Tabela 7: Teste KMO e de Bartlett.....	43
Tabela 8: Análise de componentes principais.....	43
Tabela 9: Resultado classificadores via matriz confusão.....	45
Tabela 10: Adições, exclusões e turnover da Carteira 1.....	50
Tabela 11: Análise de desempenho da Carteira 1 e do índice IBrX 100.....	52

SUMÁRIO

1. INTRODUÇÃO	12
2. REVISÃO BIBLIOGRÁFICA	14
3. METODOLOGIA	18
3.1. COLETA DOS DADOS	19
3.2. CRIAÇÃO DAS VARIÁVEIS INDEPENDENTES	21
3.3. ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)	26
3.4. TESTE KAISER-MEYER-OLKIN (KMO)	27
3.5. TESTE DE ESFERICIDADE DE BARTLETT	28
3.6. <i>SUPPORT VECTOR MACHINE</i> (SVM)	28
3.7. SVM'S NÃO LINEARES	29
3.8. REDES NEURAIS - <i>MULTILAYER PERCEPTRON</i> (MLP)	30
3.9. MLP - HIPERPARÂMETROS	32
3.10. TAXA DE APRENDIZADO	33
3.11. CAMADAS OCULTAS	34
3.12. NEURÔNIOS POR CAMADA	34
3.13. FUNÇÃO DE ATIVAÇÃO	35
3.14. FUNÇÃO DE MINIMIZAÇÃO	35
3.15. PARTIÇÃO DOS DADOS	35
3.16. MATRIZ CONFUSÃO	35
3.17. <i>BACK-TESTING</i> E AVALIAÇÃO DE DESEMPENHO DA CARTEIRA	37
3.18. SOFTWARES UTILIZADOS	39
4. DADOS	41
5. RESULTADOS	43
5.1. CLASSIFICADORES	43
5.2. CARACTERÍSTICAS DA AMOSTRA DE TESTE E DAS PREDIÇÕES	45
5.3. <i>BACK-TESTING</i>	49
6. CONCLUSÕES	56
REFERÊNCIAS	58
APÊNDICE	62

1. INTRODUÇÃO

Quando se gerencia uma carteira de ações é possível buscar diferentes objetivos de investimento: gerar retornos absolutos positivos, desempenhar melhor que um determinado índice de referência, alto pagamento de dividendos, entre outros. Cada objetivo definirá uma carteira de ações diferente, com ativos e pesos que mais reflitam o propósito final. Independentemente da finalidade esperada, o investidor deve escolher ou compor uma carteira buscando os ativos com maior potencial de retorno.

Ao longo dos anos, investidores e acadêmicos apresentaram algumas métricas e padrões que pudessem identificar ações vencedoras ao longo do tempo. Entre eles, Graham (1949), o qual pondera a seleção de ações com métricas de valor, como preço de mercado sobre lucro ou preço de mercado sobre valor de patrimônio líquido. Greenblatt (2010), por meio da metodologia *Magic Formula*, combinou métricas de valor e de qualidade das empresas para seleção das melhores ações, sugerindo o investimento naquelas que apresentassem os maiores índices de retorno sobre capital investido e de rendimento de lucros. E Novy-Marx (2013), ao demonstrar que a métrica margem bruta, ou lucro bruto sobre receita líquida, possui mais relevância para identificar ações vencedoras quando comparado as diferentes métricas de qualidade.

Neste estudo, será dado ênfase na identificação de ações que possam desempenhar melhor que o índice de referência, neste caso o IBrX 100. Como dados de entrada, usaremos informações de balanços e de demonstrativos financeiros divulgados trimestralmente pelas empresas e, para os dados de saída, o desempenho das respectivas ações nos cinquenta dias úteis subsequentes à divulgação de seus resultados. A técnica de análise de componentes principais é empregada com objetivo de diminuir a dimensionalidade da amostra, e as técnicas de classificação, no caso *support vector machine* e redes neurais, para identificação e seleção dos ativos. Tudo isso, por sua vez, visa facilitar o trabalho e a alocação de capital humano na análise das empresas, bem como, tornar mais assertivo a identificação daquelas que poderão ter um bom resultado de investimento.

Existem grandes desafios nesse tipo de abordagem. Pode-se citar que: ao se trabalhar com retornos, a memória dos dados não será considerada; a definição da amostra final pressupõe a definição de alguns filtros e, com isso, há possibilidade de incorrer em viés de sobrevivência (para se obter dados ao longo do tempo, as ações listadas em determinados períodos serão privilegiadas) e, por fim; o fato de que dado a natureza de negócio de cada empresa, as métricas fundamentalistas e de análise de empresas utilizadas para a classificação podem direcionar para interpretações e classificações incorretas.

Assim sendo, o objetivo deste trabalho é propor um modelo de seleção de ações que superem o índice IBrX 100, com base em seus registros contábeis e indicadores de mercado, a partir da utilização das técnicas de *support vector machine* e redes neurais - identificadas ao longo deste estudo como classificadores.

Este trabalho está dividido em 6 seções. Na segunda seção, são revistas as referências bibliográficas da técnica de análise de componentes principais e seu uso como técnica de pré-processamento dos dados; são também verificados estudos que utilizaram os classificadores *support vector machine* e redes neurais, verificando assim os parâmetros e resultados obtidos. Na seção 3 é apresentado em detalhes a metodologia da construção da base de dados, das técnicas utilizadas, dos critérios adotados para avaliar o desempenho da carteira formada pelos ativos selecionados e da verificação dos resultados. Na seção 4, por sua vez, é apresentado os passos da elaboração da base final. Já na seção 5, são analisados os resultados das técnicas aplicadas e o resultado da carteira composta pelas ações selecionadas. Por fim, na seção 6, são discutidas as conclusões e reflexões sobre os resultados.

2. REVISÃO BIBLIOGRÁFICA

No estudo de Novy-Marx (2014) são analisados os desempenhos de estratégias baseadas em conceitos de valor e qualidade, sendo elas: o critério de qualidade de Graham (1949), o critério de Grantham (2004), o retorno sobre o capital investido de Greenblatt (2010), os critérios de medição de lucro de Sloan (1996), a medida F-score de Piotroski (2000) e a rentabilidade bruta de Novy-Marx (2013). Para comparar os diversos critérios, as estratégias são construídas por meio do ordenamento das empresas de acordo com as respectivas métricas. Foi verificado que a métrica de rentabilidade bruta ajuda os tradicionais investidores, que priorizam métricas de mensuração de valor, a distinguir ações muito baratas de ações conhecidas como “armadilhas de valor” (*value traps*). Segundo o autor, empresas baratas e rentáveis tendem a desempenhar melhor do que empresas apenas baratas ou apenas rentáveis.

Hargraves e Mani (2015) identificam os fatores mais importantes que distingue uma ação vencedora ou, no caso, uma cesta de ações que superam o desempenho do mercado de ações australiano (ASX). Os autores utilizam a técnica de análise de componentes principais para reduzir a dimensionalidade da amostra. Foram escolhidas 22 variáveis inicialmente, de 101 ações do setor de saúde australiano no período de janeiro de 2015 a março do mesmo ano. O índice de referência para avaliação de desempenho é o ASX. Foram realizados os testes de Kaiser-Meyer-Olkin e de esfericidade de Bartlett, a fim de verificar a adequação da amostra para análise fatorial. Foram removidas as ações com baixa negociabilidade e com preço inferior a 10 centavos de dólares australianos. Como resultado, das 22 variáveis inicialmente selecionadas, quatro foram consideradas as mais importantes: i) retorno sobre investimento, ii) retorno sobre patrimônio líquido, iii) patrimônio líquido por ação e iv) receita líquida por ação.

Yang, Rea, e Rea (2016) utilizam a técnica de análise de componentes principais a fim de selecionar uma cesta reduzida de ações, que consigam replicar o comportamento do índice ASX200. Esse índice é composto pelas 200 maiores empresas em relação ao valor de mercado na bolsa de valores australiana. Para a seleção de ações foi utilizado o método de seleção variável, no qual é relacionado

com a regra de Kaiser (1960), que consiste em reter os componentes principais de uma matriz de correlação com autovalores acima de 1. O resultado demonstrou que é possível replicar, com certa margem de erro, o índice ASX200 com um menor número de ações. No entanto, verificou-se que a quantidade de ações necessárias para replicar o índice se alterou ao longo dos períodos analisados, variando de 12 ações até 25. Também foi constatado que o resultado não foi satisfatório no período da crise financeira de 2008.

Fan e Palaniswami (2001) utilizam o algoritmo *support vector machine* como método de classificação de melhores ações. Uma carteira de ações igualmente balanceada, com os ativos escolhidos pelo algoritmo obteve retorno de 208% em um período de 5 anos, significativamente maior quando comparado ao índice de referência com 71% no mesmo período. Foram extraídos os dados de ações da Bolsa de Valores Australiana no período de 1992 a 2000 e, foram considerados registros contábeis com periodicidade anual. Em seguida, foram agrupados indicadores financeiros similares em oito grupos distintos: retorno sobre capital, rentabilidade, endividamento, investimento, crescimento, liquidez de curto prazo, retorno sobre investimento e, por último, risco. Para cada ano, eram selecionadas as 25 melhores ações e classificadas como +1, e as demais -1. Dessa forma buscaram identificar as características do melhor grupo para cada ano do período.

Yu, Chen, e Zhang (2014) constroem um modelo de seleção de ações utilizando o algoritmo *support vector machine*, e tratando o problema de dimensionalidade com a técnica de análise de componentes principais. Os dados de entrada são indicadores financeiros de empresas listadas na Bolsa de Valores de Shanghai do tipo A, dos anos de 2009 a 2010. Como variável dependente, foram classificadas como +1 as ações do primeiro quartil que melhor desempenharam na amostra, e como -1 as 75% restantes. A acurácia na amostra de treinamento foi de 100% da categoria +1 e de 85% na categoria -1, enquanto na amostra de teste foi de 10% e 89%, respectivamente. O baixo desempenho na amostra de teste deve-se ao problema de *overfitting* e ao fenômeno de dados desbalanceados.

Alberg e Lipton (2017) propõem uma estratégia de investimento que constrói uma carteira de ações “hoje”, baseados em fundamentos futuros previstos. Os algoritmos

utilizados são: redes neurais - *multilayer perceptron* e redes neurais recorrentes. Os autores denotam ambos os algoritmos como *Lookahead Factor Models (LFMs)*. A construção da amostra foi baseada em uma série histórica de 5 anos de dados fundamentalistas. As simulações demonstraram que o investimento com *LFMs* baseado nos fatores previstos retornou um rendimento anualizado composto de 17,1%, comparado a 14,4% de um modelo de fator normal e, um índice de Sharpe (1966) 0,68 e 0,55, respectivamente. Para este estudo, foram consideradas todas as ações públicas negociadas com um mínimo de 12 meses consecutivos nas bolsas de valores NYSE, NASDAQ ou AMEX, no período de janeiro de 1970 a setembro 2017. Foram excluídas as ações de empresas com base fora dos EUA, empresas do setor financeiro e empresas com valor de mercado inferior a \$ 100 milhões de dólares; chegando a uma lista final de 11.815 ações. Para cada ação foram consideradas 20 variáveis em seu respectivo tempo, sendo 16 relacionadas a fundamentos das empresas e 4 relacionadas ao desempenho histórico das ações.

Maciel e Ballini (2010) analisam a aplicação de redes neurais para previsão de séries temporais em dados financeiros e, suas habilidades para prever tendências futuras nos mercados de ações dos EUA, da Europa e no Brasil. A precisão dos modelos é comparada ao método tradicional de previsão, *generalized autoregressive conditional heteroskedasticity (GARCH)*. Como variáveis, foram selecionadas os preços de fechamento dos seguintes índices: DOW JONES (EUA), S&P 500 (EUA), DAX (Alemanha), CAC 40 (França), FTSE (Inglaterra), IBEX 35 (Espanha), PSI 20 (Portugal) e IBOVESPA (Brasil), de $t-1$ até $t-4$. Os autores concluem que a aplicação de redes neurais tem a capacidade de prever a variação dos índices estudados e, se adequadamente treinadas, a robustez pode ser melhorada dependendo da estrutura da rede. Além disso, os testes Ashley-Granger-Schmalensee e Morgan-Granger-Newbold indicam que as redes neurais artificiais superam os modelos *GARCH* em termos estatísticos.

No artigo de Prado (2018) são expostos alguns dos erros mais comuns, cometidos por especialistas, ao aplicarem técnicas de aprendizado de máquina em conjuntos de dados financeiros. O motivo é que esses conjuntos exibem propriedades que violam premissas padrões em aplicações de aprendizado de máquina, incorrendo assim na produção de falsos positivos.

O trabalho de Arnott, Harvey e Markowitz (2019) propõe um protocolo para pesquisas empíricas em finanças, com aplicação em métodos estatísticos tradicionais e técnicas modernas de aprendizado de máquina. Para os autores, o escopo mais importante não é eliminar todos os falsos positivos; é enfrentar o duplo objetivo de satisfazer a minimização de estratégias falsas e não perder muitas estratégias boas, concomitantemente. O protocolo é categorizado em 7 itens, sendo eles: i) motivação da pesquisa, ii) testes múltiplos e métodos estatísticos, iii) dados e escolha da amostra, iv) validação cruzada, v) dinâmica do modelo, vi) complexidade, e por último vii) cultura de pesquisa.

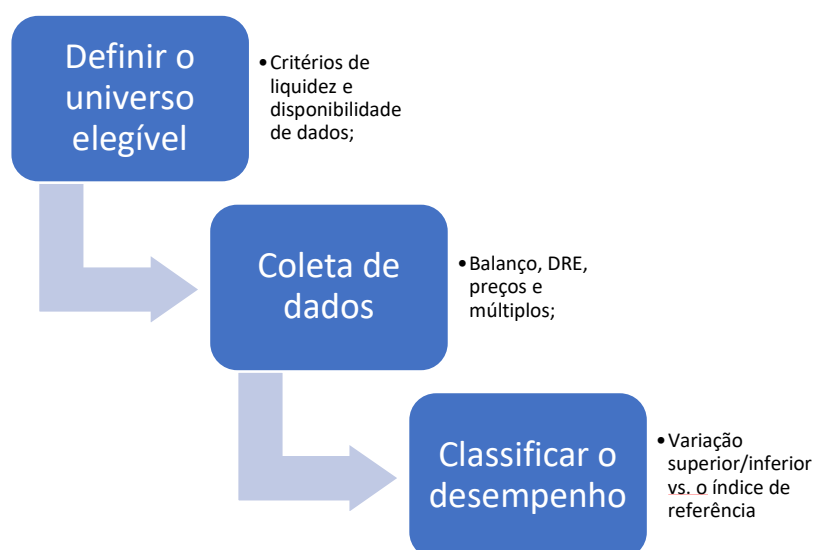
3. METODOLOGIA

Para construir a base deste estudo, são utilizados dados históricos de balanços financeiros, dos demonstrativos de resultados e dos fluxos de caixa, divulgados trimestralmente pelas empresas elegíveis, assim como preços e indicadores de mercado.

O período de extração dos dados começa em janeiro de 2001 e termina no exercício do 2º trimestre de 2019.

São considerados ativos elegíveis: ações de empresas listadas e negociadas na B3 S.A. – Brasil, Bolsa, Balcão, fusão decorrente das empresas BM&FBOVESPA e Cetip, no período citado. Como regra de elegibilidade para amostra será definido um critério de liquidez, sendo aceitas as ações que apresentaram um volume negociado acima de R\$ 3 milhões num período de três meses anteriores da divulgação de cada resultado.

Figura 1: Elaboração da base de dados



Fonte: elaboração própria

3.1. COLETA DOS DADOS

Para se obter o máximo de amostras válidas para a base final foram consideradas informações financeiras que estivessem disponíveis e, válidas ao longo dos anos, ou seja, passíveis de comparação em todo o período analisado.

Para a criação das variáveis independentes foram ponderadas as informações históricas a seguir, com base em cada divulgação de resultado em todo o período analisado. Considerar as seguintes notações:

(t): valor referente ao trimestre do exercício considerado;

(12M): valor referente aos últimos 12 meses, contados a partir do exercício considerado;

(MP12M): valor ponderado dos últimos 12 meses; e

(VF): valor de fechamento no dia útil seguinte a data base.

- Total de ativos: ativo circulante (t) mais ativo não circulante (t);
- Patrimônio líquido (t);
- Índice de liquidez corrente: ativo circulante (t) sobre passivo circulante (t). Mede a capacidade de pagamento de uma empresa no curto prazo;
- Dívida líquida: total das dívidas de curto e de longo prazo (t) menos as disponibilidades circulantes (t);
- Margem bruta: lucro bruto (t) sobre a receita líquida (t);
- LAJIDA (t): lucro antes dos juros, impostos, depreciação e amortização;
- Dívida líquida sobre LAJIDA: este múltiplo é característico para mensurar o nível de endividamento da empresa, extraída pelo quociente das linhas de dívida líquida (dívida bruta menos a disponíveis no circulante) (t) sobre o LAJIDA (12M);
- LPA - lucro líquido por ação: lucro líquido (t) sobre quantidade de ações emitidas (MP12M);
- ROIC - retorno sobre o capital investido: lucro operacional líquido após os impostos (12M) sobre o valor contábil do capital investido (t);
- ROA - retorno sobre o ativo: lucro líquido (12M) sobre o total de ativos (t);

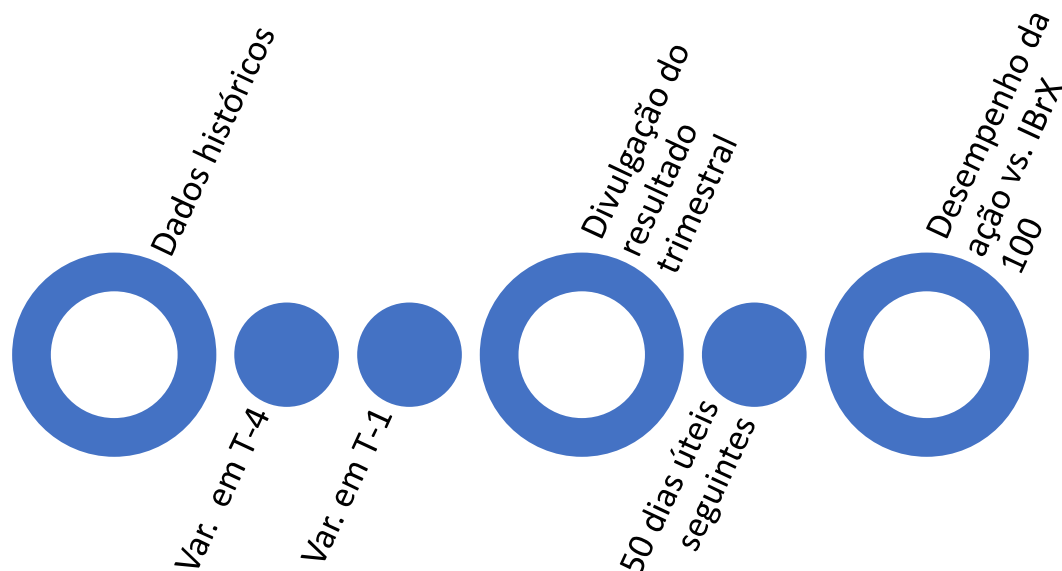
- Preço sobre valor patrimonial da ação: preço por ação (VF) sobre valor do patrimônio líquido (t), por ação emitida;
- Rendimento do dividendo: valor dos dividendos provisionados (12M) sobre o valor de mercado da companhia (VF);
- Beta: do ativo sobre o Ibovespa no período de 2 anos (VF);
- Valor de empresa sobre receita líquida: este múltiplo é uma referência de valor da ação, ou seja, se o preço de mercado está caro ou barato. O múltiplo é o resultado do quociente do valor da empresa (VF) sobre o valor de receita líquida (12M);
- Preço sobre vendas: outro múltiplo de referência de mensuração de valor da ação. É o resultado do quociente do valor de mercado da empresa (VF) sobre a receita líquida (12M); e
- Variação de preço da ação nos períodos: 5 dias (VF), 1 mês (VF) e 2 meses (VF).

A data base para cada amostra é a data da divulgação do resultado trimestral. Para os indicadores extraídos do mercado, como o preço da ação ou múltiplos de valor, são considerados os valores de fechamento do dia útil seguinte a data base, com objetivo de padronizar os efeitos gerados devido a divulgação de resultados. As empresas divulgam seus balanços antes da abertura do mercado ou após o seu fechamento. Logo, os preços das ações de empresas que divulgaram antes refletem os impactos das novas informações no mesmo dia, já as demais ações são impactadas no dia seguinte de negociação.

A variável dependente é binária, ou seja, assume o valor 0 ou 1. Caso a ação da amostra tenha desempenhado melhor que o índice de referência, nos 50 dias úteis subsequentes da data base, atribui-se o valor de 1 e, se for igual ou pior, assume o valor 0. O prazo de 50 dias úteis é adotado com intuito de isolar o efeito da próxima divulgação de resultado, uma vez que o tempo médio entre cada divulgação são 90 dias corridos. O índice de referência considerado é o IBrX 100, calculado pela B3, que tem como objetivo ser o indicador do desempenho médio das cotações dos 100 ativos de maior negociabilidade e representatividade do mercado de ações brasileiro. Devido

a sua composição e a sua regra de ponderação, entende-se que este índice é menos concentrado setorialmente e representa uma fração maior do mercado.

Figura 2: Janela de estudo para cada ativo



Fonte: elaboração própria

3.2. CRIAÇÃO DAS VARIÁVEIS INDEPENDENTES

A partir dos dados dos ativos elegíveis coletados, são criadas as seguintes variáveis independentes para a preparação da base final, sendo: 8 que refletem a rentabilidade da empresa, 4 variáveis binárias, 5 de liquidez e endividamento, 7 de mensuração de valor e 4 que refletem o comportamento da ação no mercado; totalizando 28 variáveis. Para as descrições das variáveis, considerar as seguintes notações:

- (t): referente ao exercício à data base;
- (t-1): referente ao exercício anterior à data base; e
- (t-4): referente a 4 exercícios anteriores à data base.

Tabela 1: Variáveis de rentabilidade

Variável	#	Média	Erro padrão	Mediana	Modo	Desvio padrão	Variância da amostra	Mínimo	Máximo	Nível de confiança (95,0%)
<i>GM_Y</i>	1	-0,41	0,14	-0,22	0	8,14	66,18	-66,92	64,46	0,27
<i>GM_Q</i>	2	-0,06	0,13	-0,02	1,23	7,63	58,18	-67,53	72,55	0,26
<i>ROA_Y</i>	3	-0,22	0,08	-0,16	-5,67	4,77	22,78	-50,33	23,92	0,16
<i>ROA_Q</i>	4	-0,12	0,03	-0,05	-0,62	1,92	3,69	-37,64	16,94	0,06
<i>ROA_N</i>	5	7,37	0,12	6,45	11,08	6,86	47,07	-40,7	40,42	0,23
<i>ROIC_Y</i>	6	-0,08	0,14	-0,09	-1,62	8,35	69,67	-60,13	85,85	0,28
<i>ROIC_Q</i>	7	-0,19	0,07	-0,05	1,32	4,12	17	-56,87	57,55	0,14
<i>ROIC_N</i>	8	16,85	0,33	13,1	9,78	19,41	376,66	-16,19	223,96	0,65

Fonte: elaboração própria

Sendo elas:

- *GM_Y*: a diferença da margem bruta em (t) contra (t-4);
- *GM_Q*: a diferença da margem bruta em (t) contra (t-1);
- *ROA_Y*: a diferença do múltiplo ROA em (t) contra (t-4);
- *ROA_Q*: a diferença do múltiplo ROA em (t) contra (t-1);
- *ROA_N*: o valor em nível do múltiplo ROA em (t);
- *ROIC_Y*: a diferença do múltiplo ROIC em (t) contra (t-4);
- *ROIC_Q*: a diferença do múltiplo ROIC em (t) contra (t-1); e
- *ROIC_N*: o valor em nível do múltiplo ROIC em (t).

Analisando as variáveis de rentabilidade, nota-se que todas apresentam números negativos, exceto as que refletem o valor em nível. Com isso, é possível induzir que as empresas contidas na amostra não apresentaram crescimento de rentabilidade na média. Nota-se também, que o múltiplo ROA é mais volátil que o ROIC no período, apresentando um desvio padrão de 6,86 e 19,41, respectivamente.

Tabela 2: Variáveis binárias

Variável	#	Média	Erro padrão	Mediana	Modo	Desvio padrão	Variância da amostra	Mínimo	Máximo	Nível de confiança (95,0%)
<i>EBITDA_Y</i>	1	0,4	0,02	1	1	0,92	0,84	-1	1	0,03
<i>EBITDA_Q</i>	2	0,32	0,02	1	1	0,95	0,9	-1	1	0,03
<i>EPS_Y</i>	3	0,13	0,02	1	1	0,99	0,98	-1	1	0,03
<i>EPS_Q</i>	4	0,04	0,02	1	1	1	1	-1	1	0,03

Fonte: elaboração própria

Sendo elas:

- *EBITDA_Y*: assume o valor de 1 se o LAJIDA (trimestral) apresentado em (t) variou positivamente versus (t-4), e assume o valor de -1 nos demais resultados;
- *EBITDA_Q*: assume o valor de 1 se o LAJIDA (trimestral) apresentado em (t) variou positivamente versus (t-1), e assume o valor de -1 nos demais resultados;
- *EPS_Y*: assume o valor de 1 se o lucro por ação (trimestral) apresentado em (t) variou positivamente versus (t-4), e assume o valor de -1 nos demais resultados; e
- *EPS_Q*: assume o valor de 1 se o lucro por ação (trimestral) apresentado em (t) variou positivamente versus (t-1), e assume o valor de -1 nos demais resultados.

Tabela 3: Variáveis de liquidez e endividamento

Variável	#	Média	Erro padrão	Mediana	Modo	Desvio padrão	Variância da amostra	Mínimo	Máximo	Nível de confiança (95,0%)
<i>NETDEBTEBITDA_Y</i>	1	0,08	0,12	0,01	-4,09	7,09	50,3	-223,02	189,52	0,24
<i>NETDEBTEBITDA_Q</i>	2	-0,03	0,11	0,01	0,28	6,26	39,13	-238,71	184,77	0,21
<i>NETDEBTEBITDA_N</i>	3	1,25	0,09	1,31	-3,31	5,22	27,26	-220,19	65,51	0,18
<i>CURRATIO_Y</i>	4	0,01	0,01	0	-0,02	0,74	0,55	-13,58	5,32	0,02
<i>CURRATIO_Q</i>	5	0,01	0,01	0,01	0,16	0,49	0,24	-5,45	5,52	0,02

Fonte: elaboração própria

Sendo elas:

- *NETDEBTEBITDA_Y*: a diferença do múltiplo Dívida líquida sobre LAJIDA em (t) contra (t-4);
- *NETDEBTEBITDA_Q*: a diferença do múltiplo Dívida líquida sobre LAJIDA em (t) contra (t-1);
- *NETDEBTEBITDA_N*: o valor em nível do múltiplo em (t);
- *CURRATIO_Y*: a diferença do múltiplo Índice de liquidez corrente em (t) contra (t-4); e
- *CURRATIO_Q*: a diferença do múltiplo Índice de liquidez corrente em (t) contra (t-1).

O nível de endividamento médio da amostra é de 1,25 segundo a variável dívida líquida sobre LAJIDA. As variáveis de liquidez corrente apresentam menores valores de desvio padrão, ao serem comparadas com as variáveis do múltiplo dívida líquida sobre LAJIDA. Pode-se induzir que a volatilidade do investimento em capital de giro das empresas na amostra é menor que os dados de dívida bruta e LAJIDA. Isso pode ser explicado devido ao fato de muitas empresas da amostra possuírem parte da dívida ou, em sua totalidade em moeda estrangeira.

Tabela 4: Variáveis de mensuração de valor

Variável	#	Média	Erro padrão	Mediana	Modo	Desvio padrão	Variância da amostra	Mínimo	Máximo	Nível de confiança (95,0%)
<i>P/B_Z</i>	1	0,18	0,02	0,22	-0,58	1,14	1,3	-2,65	2,74	0,04
<i>P/B_N</i>	2	2,98	0,07	1,68	1,16	3,98	15,85	0,01	44,27	0,13
<i>EV_SALES_Z</i>	3	0,08	0,02	0,05	1,21	1,15	1,32	-2,74	2,78	0,04
<i>EV_SALES_N</i>	4	3,32	0,39	1,91	1,1	22,58	509,78	0,12	628,91	0,76
<i>P_SALES_Z</i>	5	0,18	0,02	0,27	1,3	1,18	1,38	-2,9	2,82	0,04
<i>P_SALES_N</i>	6	1,72	0,03	1,23	0,48	2,03	4,1	0	34,06	0,07
<i>DVD_YIELD</i>	7	7,67	0,68	3,71	1,27	39,68	1.574,61	0	1.037,82	1,33

Fonte: elaboração própria

Sendo elas:

- *P/B_Z*: desvios padrões da média do múltiplo P/VPA em (t), com relação ao histórico dos 12 trimestres anteriores;
- *P/B_N*: o valor em nível do múltiplo P/VPA em (t);
- *EV_SALES_Z*: desvios padrões da média do múltiplo Valor de empresa sobre receita líquida em (t), com relação ao múltiplo histórico dos 12 trimestres anteriores;
- *EV_SALES_N*: o valor em nível do múltiplo Valor de empresa sobre receita líquida em (t);
- *P_SALES_Z*: desvios padrões da média do múltiplo Preço sobre receita líquida em (t), com relação ao múltiplo histórico dos 12 trimestres anteriores;
- *P_SALES_N*: o valor em nível do múltiplo preço sobre vendas em (t); e
- *DVD_YIELD*: valor do índice Rendimento do dividendo em (t).

Ao analisarmos as variáveis de mensuração de valor, os múltiplos *P/B_N*, *EV_SALES_N* e *P_SALES_N* apresentaram os valores médios 2,98, 3,32 e 1,72, respectivamente. Observa-se também que as variáveis com maiores valores de desvio padrão são: valor de empresa sobre receita líquida e o rendimento sobre dividendo.

Tabela 5: Variáveis de desempenho de mercado

Variável	#	Média	Erro padrão	Mediana	Modo	Desvio padrão	Variância da amostra	Mínimo	Máximo	Nível de confiança (95,0%)
<i>VARPRICE_2M</i>	1	4,03	0,29	3,14	0,00	17,13	293,46	-65,06	173,71	0,58
<i>VARPRICE_1M</i>	2	2,24	0,18	2,01	0,00	10,77	115,93	-55,89	62,48	0,36
<i>VARPRICE_5D</i>	3	0,24	0,09	0,26	0,00	5,36	28,77	-33,33	37,66	0,18
<i>RAW_BETA</i>	4	0,70	0,01	0,67	0,43	0,30	0,09	-0,25	1,83	0,01

Fonte: elaboração própria

Sendo elas:

- *VARPRICE_2M*: variação do preço da ação nos últimos 2 meses (ajustado por todos os eventos relacionados ao ativo), com base no preço do dia seguinte da data base;
- *VARPRICE_1M*: variação do preço da ação nos últimos 30 dias (ajustado por todos os eventos relacionados ao ativo), com base no preço do dia seguinte da data base;
- *VARPRICE_5*: variação do preço da ação nos últimos 5 dias (ajustado por todos os eventos relacionados ao ativo), com base no preço do dia seguinte da data base; e
- *RAW_BETA*: valor do índice beta em (t).

Referente as variações históricas, é possível verificar que quanto maior o prazo maior é a dispersão e maior os desvios dos retornos. O beta de todos os ativos da amostra apresentou valor médio de 0,70.

3.3. ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)

Devido à grande quantidade de variáveis inicialmente consideradas, incorreremos na Maldição da Dimensionalidade. Este termo foi empregado por Bellman (1961) e se refere à existência de uma relação direta entre dimensionalidade dos dados, e a quantidade de dados necessária para possibilitar o sucesso da tarefa de aproximação. A Maldição da Dimensionalidade encontra-se na limitação de se obter um número significativo de amostras, que preencham o volume da estrutura de dados, e que permita uma estimação adequada da superfície. A ausência de amostras de dados em certas regiões fará com que os estimadores preencham os espaços de forma aleatória, degenerando a estrutura de dados original.

A fim de contornar o problema acima é aplicado a técnica de análise de componentes principais (ACP). De acordo com Kitani (2013), é a mais antiga, conhecida e utilizada ferramenta de exploração de dados em alta dimensão. Foi desenvolvida em 1901 por Karl Pearson para aplicações em ambientes biológicos utilizando uma abordagem geométrica. A técnica tem sido aplicada em diversos tipos de estudos como parte do pré-processamento dos dados e, também para atacar o problema da

dimensionalidade (MBELEDUGU; ODOH; UMEH, 2012). A ACP é um método não supervisionado de descobrir as características mais relevantes dos dados.

A ideia central da análise de componentes principais é reduzir a dimensionalidade de um conjunto de dados, mantendo o máximo possível à variância presente no conjunto de dados original. Isso é alcançado mediante um procedimento matemático, que utiliza uma transformação ortogonal, para converter um conjunto de observações de variáveis possivelmente correlacionadas em um conjunto de valores de variáveis linearmente não correlacionados, chamados de componentes principais (PCs). Em outras palavras, transforma a base original em um novo conjunto de variáveis que não estão correlacionados, e que são ordenados para que os primeiros retenham a maior parte da variação presente em todos as variáveis originais (JOLLIFFE, 2002). O número de componentes principais é sempre menor ou igual ao número de variáveis originais.

A ACP trabalha com a matriz de covariância das amostras procurando uma nova base ortogonal, que minimize o erro de reconstrução de dados de entrada e maximize a variância dessa nova base. Dessa forma, a ACP permite representar os dados de entrada como combinações lineares das suas próprias projeções na nova base. (KITANI, 2013)

3.4. TESTE KAISER-MEYER-OLKIN (KMO)

A fim de verificar a adequação da amostra para a análise fatorial, será utilizado a estatística Kaiser-Meyer-Olkin (KMO). O teste KMO é um critério para identificar se um modelo de análise fatorial, que está sendo utilizado, é adequadamente ajustado aos dados, testando sua consistência geral. Esse método verifica se a matriz de correlação inversa é próxima da matriz diagonal. Basicamente, consiste em comparar os valores dos coeficientes de correlação linear, observando os valores dos coeficientes de correlação parcial.

Figura 3: Kaiser-Meyer-Olkin (KMO)

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2}$$

Onde:

i : referente a linha da matriz;

j : referente a coluna da matriz;

r_{ij} : coeficiente de correlação entre variáveis; e,

a_{ij} : coeficiente de correlação parcial.

O valor de KMO próximo de 0 indica que a análise fatorial pode não ser adequada, pois possui uma correlação fraca entre as variáveis. Quanto mais próximo de 1, mais adequada é a utilização. Para que a amostra seja adequada, espera-se que o resultado seja acima de 0,6. (HARGRAVES; MANI, 2015)

3.5. TESTE DE ESFERICIDADE DE BARTLETT

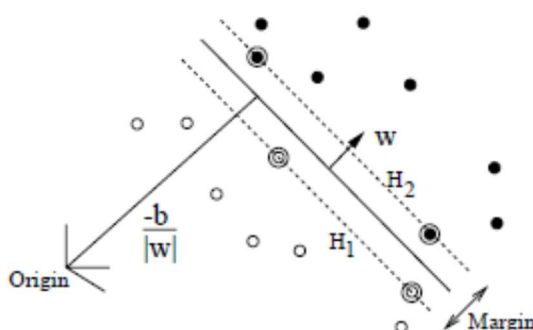
Juntamente com o teste KMO, será realizado o teste de esfericidade de Bartlett, que testa a hipótese de que as variáveis não sejam correlacionadas na população. Este teste verifica a hipótese de a matriz de correlação populacional ser igual à matriz identidade, ou seja, admitir a ausência de associação linear entre as variáveis estudadas. O resultado esperado é que seja significativo, assim rejeitando a hipótese nula.

3.6. SUPPORT VECTOR MACHINE (SVM)

O *support vector machine* (SVM) é uma técnica de aprendizado de máquina que busca a minimização estrutural do risco, usando um mapeamento, para criar um hiperplano ótimo no espaço das características para separar linearmente duas classes distintas. Um modelo de classificação SVM requer dois tipos componentes básicos: os vetores

de suporte em si e o limite ideal de decisão. É considerada uma técnica linear e supervisionada. Contudo, se usarmos um kernel não linear podemos resolver classificações não lineares também. Fundamentalmente, o SVM olha para os representantes das fronteiras entre duas classes distintas e, tenta maximizar a distância entre vetores de suporte, que são criados a partir desses representantes. Após a etapa de aprendizado, o SVM é capaz de classificar qualquer nova amostra.

Figura 4: Exemplo de separador linear do hiperplano



Fonte: Burges (1998)

Vemos que na figura acima, as amostras marcadas com um círculo são os vetores de suporte de cada classe. O SVM busca maximizar a distância do hiperplano perpendicular à W que o mantenha afastado das margens.

Os vetores de suporte são um pequeno subconjunto de dados de treinamento que são extraídos pelo algoritmo. Dependendo de como o kernel do produto interno é construído, podemos ter diferentes máquinas de aprendizagem caracterizadas por superfícies de decisão não-lineares. (HAYKIN, 1999)

3.7. SVM'S NÃO LINEARES

Segundo Lorena e Carvalho (2007), o SVM é eficaz na classificação de conjuntos de dados linearmente separáveis ou que possuem uma distribuição aproximadamente linear. Porém, há muitos casos em que não é possível dividir satisfatoriamente os

dados de treinamento por um hiperplano. Neste caso, é possível a utilização de funções kernel na não-linearização do SVM tornando o algoritmo mais eficiente, pois permite a construção de simples hiperplanos em um espaço de alta dimensão de forma tratável do ponto de vista computacional.

Suponhamos um mapeamento de dados: $\phi : \chi \rightarrow K \subseteq \ell_2^n$. Um kernel K é uma função do produto escalar: $\chi \times \chi \rightarrow \mathbb{R}$ em K , por exemplo, para todos $x_i, x_j \in \chi$,

$$K(x_i, x_j) \stackrel{\text{def}}{=} \langle \phi(x_i), \phi(x_j) \rangle = \langle x_i, x_j \rangle. \text{ (HERBRICH, 2002)}$$

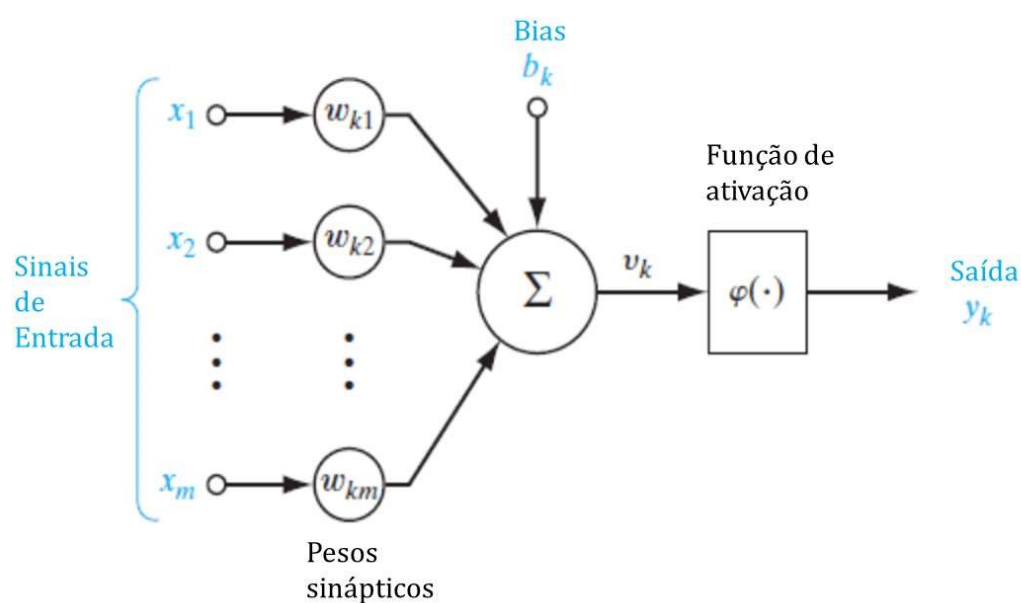
Devido a dinâmica das séries temporais financeiras terem uma característica não linear (QI; MADDALA, 1999), pode-se induzir que a utilização de funções kernel não lineares poderiam apresentar resultados melhores quando comparável a função linear. Os kernel não lineares mais conhecidas são os polinomiais e os Gaussianos, também conhecidos como função de base radial (*RBF*). Segundo o trabalho de Cao e Tay (2003), o tempo de treinamento é menor e os resultados obtidos são superiores utilizando o kernel Gaussiano, quando comparado ao kernel polinomial. Portanto, para este trabalho e, para nível de comparação de desempenho dos classificadores, serão utilizados o kernel linear e o kernel Gaussiano.

3.8. REDES NEURAIS - *MULTILAYER PERCEPTRON (MLP)*

O trabalho com redes neurais artificiais, comumente referido como “redes neurais”, tem sido motivado desde o início pelo reconhecimento de que o cérebro humano calcula de uma maneira totalmente diferente do computador digital convencional. O cérebro tem a capacidade de organizar seus componentes estruturais, conhecidos como neurônios, para executar certos cálculos muitas vezes mais rápido que um computador convencional. Na sua forma mais geral, uma rede neural é um algoritmo projetado para modelar a forma que o cérebro desempenha determinada tarefa ou função de interesse; a rede é geralmente implementada mediante o uso de componentes eletrônicos ou é simulada em um programa de computador. Um neurônio é uma unidade de processamento básico em qualquer rede neural.

O *perceptron* é um tipo de rede neural artificial inventada por Frank Rosenblatt em 1958 no artigo “*The Perceptron: a probabilistic model for information storage and organization in the brain*”, e pode ser visto como o tipo mais simples de rede neural *feed forward*. O *perceptron* torna um vetor de entrada $X \in \mathbb{R}^n$ e, calcula a combinação linear dessas entradas com o vetor de pesos $W \in \mathbb{R}^n$, e coloca na saída o valor de 1 se o resultado é maior dado algum limite e, -1 se for menor. A principal limitação desta rede é que é um classificador linear.

Figura 5: Representação clássica de um modelo de neurônio artificial



Fonte: Adaptado de Haykin (1999)

Sendo:

x_1, \dots, x_m : são as entradas;

w_{k1}, \dots, w_{km} : são os pesos sinápticos de cada entrada;

Σ : é a unidade somadora;

b_k : tem a função de aumentar ou reduzir o sinal de saída;

v_k : conexão sináptica vindo de outro neurônio; e

$\varphi(\cdot)$: função de ativação que dispara a saída caso um sinal y_k .

A função de ativação define o valor de saída de um neurônio a partir do sinal de entrada. Ou seja, é a transformação que fazemos ao longo do sinal de entrada para então ser enviada a próxima camada de neurônios. Segundo Haykin (1999), existem 3 tipos básicos de funções: i) função de ativação binária, ii) função de ativação linear e iii) função de ativação sigmoidal ou logística. Esta última é a mais usual das curvas e o grande benefício dessa em relação à linear ou binária, é ser diferenciável.

O *perceptron* sozinho não consegue tratar problemas não lineares. Contudo, um *perceptron* que possui múltiplas camadas pode dividir o espaço de entrada de maneiras mais complexas e não lineares, resolvendo a limitação da rede linear de uma camada. Logo, uma rede neural que é composta por uma série de unidades sensoriais que constituem a camada de entrada, uma ou mais camadas ocultas de nós e, uma camada de saída é comumente chamado como *multilayer perceptron* (*MLP*), que representa uma generalização do *perceptron*. O sinal de entrada propaga para frente através da rede, passando por cada camada (conhecido como processo de *feed forward*). O *MLP* vem sendo aplicado com sucesso para resolver problemas mais difíceis a partir do treinamento supervisionado, juntamente com outro algoritmo conhecido como *backpropagation error* e, a base principal desse algoritmo é o gradiente descendente. O treinamento de uma rede *MLP* via *backpropagation error* envolve 3 estágios distintos: i) *feed forward* no qual os sinais de entrada são propagados através da rede, seguindo os produtos internos e funções de ativação, ii) determinação do erro nas saídas e, posterior retro propagação do erro e iii) atualização dos pesos e retorno ao primeiro estágio.

3.9. *MLP* - HIPERPARÂMETROS

Uma *MLP* tem vários hiperparâmetros que afetam o comportamento da rede. Fundamentalmente, o erro produzido por uma *MLP* é função dos parâmetros livres. Os principais hiperparâmetros são:

- taxa de aprendizagem η - é o tamanho do ajuste dos pesos a cada interação.

- o número de neurônios por camada oculta também é ajustável. Apesar de sua importância, não há uma fórmula decisiva para definir o número correto de neurônios e, em decorrência disso, acaba por ser definido por meio de experimentação. A escolha aleatória da quantidade aplicada pode implicar em problemas de *underfitting* ou *overfitting*, bem como aumentar a complexidade e o tempo de treinamento sem necessidade. No trabalho de Sheela e Deepa (2013) são revisados os métodos de determinação do número de neurônios em redes neurais nos últimos 20 anos e, são listados 9 modelos com seus desempenhos medidos por meio do erro quadrático médio. Para este trabalho, será seguido algumas heurísticas para determinar a quantidade de neurônios numa rede.

Tabela 6: Quantidade de neurônios nas camadas

Definição do número de neurônios
$N_h = 2n + 1$
$N_h = n + D$
$N_h = (n + D) / 2$
$N_h = D$

Onde:

N_h : quantidade de neurônios;

n : a dimensionalidade das variáveis independentes; e

D : o número de classe.

- o número de camadas ocultas depende da complexidade do problema. Como cada neurônio cria um hiperplano de decisão, uma maior quantidade de neurônios aumenta a fragmentação do espaço em regiões convexas. Quanto maior o número de camadas ocultas, maior será a abstração do formato da segmentação do espaço de entrada.

3.10. TAXA DE APRENDIZADO

Para este trabalho, será utilizado o otimizador Adam. Inicialmente publicado na conferência de ICLR 2015, e desenvolvido por Diederik P. Kingman e Jimmy Lei Ba, é um método de taxa de aprendizado adaptável. Ou seja, calcula taxas de aprendizado individuais para diferentes parâmetros, e usa estimativas do primeiro e do segundo momento do gradiente para adaptar a taxa de aprendizado para cada peso da rede neural. Ele foi escolhido por ser considerado computacionalmente eficiente. Possui poucos requisitos de memória, é invariável ao redimensionamento diagonal dos gradientes e é adequado para problemas grandes em termos de dados e / ou parâmetros. Foram mantidos os parâmetros padrões do modelo sendo:

- Taxa de aprendizado = 0.001
- Beta 1 - a taxa de decaimento exponencial para o 1º momento estima = 0.9
- Beta 2 - a taxa de decaimento exponencial para o 2º momento estima = 0.999

3.11. CAMADAS OCULTAS

Como descrito acima, há inúmeras maneiras de preparar e treinar a rede neural. O número de neurônios de entrada é o mais direto, uma vez que deve conter o mesmo número de variáveis. No caso deste estudo, será a quantidade de fatores do ACP. Já para o número de camadas ocultas faremos a análise utilizando 1 e 2 camadas; como já dito anteriormente, quanto maior a quantidade de camadas, maior a complexidade e fragmentação da rede.

3.12. NEURÔNIOS POR CAMADA

Já para o número de neurônios, depende da quantidade de componentes que melhor representarão a variância das variáveis independentes, sempre com o objetivo da redução de dimensionalidade. Será utilizada as fórmulas da Tabela 2 para definição dos neurônios; logo, serão testadas 3 configurações de redes neurais sendo todas com 2 neurônios na saída (devido o resultado ser binário).

3.13. FUNÇÃO DE ATIVAÇÃO

Por se tratar de um problema de classificação, será usada nas camadas ocultas a função de ativação sigmoidal devido a suas características de não linearidade, ser diferenciável e assumir valores entre 0 e 1. Na camada de saída, será usada a função *softmax* sendo essa um tipo de função sigmoidal com valores de saída entre 0 e 1. Porém, a característica principal da função *softmax* é que os valores são divididos pela soma das saídas, assim, retornando à probabilidade da entrada estar em uma determinada classe.

3.14. FUNÇÃO DE MINIMIZAÇÃO

As funções de minimização mais comumente usadas são: *mean square error (mse)*, *cross entropy* e *binary cross entropy*. Por se tratar de uma saída binária, será utilizada a *binary cross entropy*.

3.15. PARTIÇÃO DOS DADOS

Para treinamento e validação dos classificadores, será empregada uma prática comum que é a partição dos dados. Geralmente é definida uma amostra maior para treinamento, a fim de identificar todos os padrões históricos dos dados e, uma amostra menor e mais recente para conferência e validação dos algoritmos. Para este estudo, usaremos 80% dos dados da amostra final para treinamento e os 20% para amostra de teste.

3.16. MATRIZ CONFUSÃO

Para verificação da eficácia dos classificadores na amostra de teste será utilizado a matriz confusão, sendo essa uma tabela que mostra as frequências de classificação para os valores previstos e para os reais.

Figura 6: Matriz confusão

		Valores previstos	
		0	1
Valores reais	0	Verdadeiro negativo (VN)	Falso positivo (FP)
	1	Falso negativo (FN)	Verdadeiro positivo (VP)

Sendo:

- Verdadeiro negativo (VN) e verdadeiro positivo (VP): ocorre quando no conjunto real dos dados as classes foram previstas corretamente; e
- Falso positivo (FP) e falso negativo (FN): ocorre quando no conjunto real dos dados as classes foram previstas incorretamente.

Podem ser extraídos alguns conceitos decorrentes da matriz:

- Acurácia: proporção global do acerto;

$$\text{acurácia} = (VN + VP) / (VN + VP + FP + FN)$$
- Precisão: proporção de acerto das previsões positivas;

$$\text{precisão} = (VP) / (VP + FP)$$
- Sensitividade (*Recall*): proporção de acertos nas observações positivas;

$$\text{sensitividade} = VP / (VP + FN)$$
- Especificidade: proporção de acerto das previsões negativas;

$$\text{especificidade} = VN / (VN + FN)$$
- *F1-Score*: média harmônica entre precisão e sensibilidade;

$$F1\text{-score} = 2*VP / (2*VP + FP + FN)$$

Vale pontuar que, dependendo do que se busca classificar, alguns conceitos tornam-se mais importantes que outros. Para este estudo, trata-se do classificador com maior acerto global, ou seja, com maior acurácia.

3.17. BACK-TESTING E AVALIAÇÃO DE DESEMPENHO DA CARTEIRA

A partir da verificação do classificador com maior acurácia é efetuado o procedimento de *back-testing* de uma carteira teórica, representada pelo nome Carteira 1, das ações selecionadas por esse classificador na amostra de teste. O procedimento e elaboração da carteira seguem os seguintes critérios:

- Carteira composta com ações selecionadas como 1 do classificador com maior acurácia;
- Os ativos que compõem a carteira terão o mesmo peso (igualmente balanceada), somando 100%. Não serão consideradas posições vendidas;
- As datas de compra e de venda das ações selecionadas seguirão o mesmo critério utilizado para a definição da variável dependente, ou seja, a compra será efetuada no primeiro dia útil seguinte da divulgação do resultado e a data de venda será após 50 dias úteis da data de compra;
- Para cada inclusão ou exclusão de ativo, as ações que compõem a carteira são rebalanceadas, com todos os ativos somando 100%;
- Os preços considerados para as transações - inclusão, exclusão e rebalanceamento da carteira - são o de fechamento do dia, com intuito de evitar as variações dos ativos ao longo do dia;
- Para os dias (ou períodos) que a carteira não possuir ações, serão considerados 100% das variações do IBrX 100 como desempenho da Carteira 1;
- O desempenho da carteira será comparado com o índice IBrX 100;
- Todos os preços considerados para a avaliação do desempenho e, para a movimentação da carteira, estão ajustados por todos os eventos ocorridos no período de análise; e
- Não são considerados custos de transação.

O retorno da carteira é calculado diariamente com base na somatória retornos de cada ativo ponderado pela sua participação. Algebricamente, temos:

$$R_C = \sum_{i=1}^n w_{i,t-1} (\ln(p_{i,t}) - \ln(p_{i,t-1}))$$

$$w_{i,t-1} = \frac{1}{A_{t-1}}$$

Onde:

R_C = retorno da carteira em t ponderado pelo peso w em $t-1$;

$w_{i,t}$ = participação do ativo i no tempo t ;

$p_{i,t}$ = preço de fechamento das ações da empresa i no tempo t ;

A_t = quantidade de ativos na carteira no tempo t .

O risco da carteira é analisado pelas métricas de volatilidade (σ) e de risco ativo da carteira (TE_C). A volatilidade será calculada por meio dos retornos absolutos diários, em um prazo de 66 dias úteis (equivalente a 90 dias corridos) e, demonstrada em formato anualizado. O risco ativo, também conhecido como *tracking error*, mede a divergência entre o comportamento de preço de um ativo, ou de uma carteira de ações, com relação a um índice de referência. O risco ativo será calculado por meio dos retornos absolutos diários, em um prazo de 252 dias úteis (equivalente a 12 meses) e, demonstrado também em formato anualizado. Ambos são definidos algebricamente como:

$$\sigma_i = \left(\sqrt{\frac{\sum_{i=1}^n (r_i - \bar{r})^2}{n - 1}} \right) * \sqrt{252}$$

Onde:

σ_i = volatilidade anualizada no período i ;

r_i = retorno no período i ;

\bar{r}_i = retorno médio;

n = número de observações no período i .

$$TE_{C,i} = \left(\sqrt{\frac{\sum_{i=1}^n (r_{C,i} - r_{B,i})^2}{n - 1}} \right) * \sqrt{252}$$

Onde:

$TE_{C,i}$ = risco ativo da carteira no período i ;

$r_{C,i}$ = retorno da carteira no período i ;

$r_{B,i}$ = retorno do índice de referência no período i ;

n = número de observações no período i .

A fim de verificar se o desempenho da carteira formada pelas ações selecionadas supera o retorno do índice de referência, levando em consideração o seu risco ativo, é calculado o *information ratio* (IRC) desenvolvido por Treynor e Black (1973). Além disso, é uma variação de Sharpe (1966) como métrica de avaliação de risco-retorno. O *information ratio* é apresentado em formato anualizado e é definido como:

$$IRC = \frac{E [r_{C,i} - r_{B,i}]}{TE_{C,i}}$$

Onde:

IRC_{i} = *information ratio* da carteira no período i ;

$r_{C,i}$ = retorno da carteira no período i ;

$r_{B,i}$ = retorno do índice de referência no período i ;

$TE_{C,i}$ = desvio padrão do excesso de retorno da carteira sobre o índice de referência, ou seja, o risco ativo.

3.18. SOFTWARES UTILIZADOS

Para extração dos dados foi utilizado o programa Bloomberg. Para preparação da base de dados e do procedimento de *back-testing* foi utilizado o Excel. As demais técnicas deste estudo foram implementadas no software Jupyter Notebook versão 5.7.8, que utiliza a linguagem de programação Python. Nas técnicas de ACP e SVM, utilizou-se a biblioteca de código aberto *scikit-learn*. Já a técnica *MLP* utilizou-se a biblioteca de rede neural de código aberto Keras.

4. DADOS

Todas as informações para a criação da base de dados foram coletadas da ferramenta Bloomberg. Inicialmente, extraíram-se os dados de todas as empresas listadas na B3, que divulgaram seus balanços trimestrais entre janeiro de 2001 até o exercício do segundo trimestre de 2019, contabilizando aproximadamente 8 (oito) mil amostras contendo: o código de negociação da empresa na bolsa, a data de divulgação de resultado e o exercício referente a essa divulgação. O primeiro tratamento dos dados foi a filtragem de ações a partir do critério de negociabilidade descrita na seção **3. Metodologia** (média diária acima de R\$ 3 milhões nos três meses anteriores que antecederam a divulgação de resultado), reduzindo o número de amostras para 5.014. Após essa primeira filtragem, foram coletados os dados históricos de cada amostra, considerando como data base a data de divulgação do balanço.

Como segundo filtro, foram consideradas algumas regras utilizadas por Almeida, Campello, Laranjeira, e Weisbenner (2009): retirar da amostra empresas com o valor total de ativos abaixo de USD 10 milhões - neste trabalho usaremos o limite de R\$ 30 milhões - e retirar empresas que apresentaram crescimento de total de ativos maior que 100% em um período de 12 meses, a fim de retirar as ações que sofreram fusões e/ou aquisições. Estes critérios reduziram o tamanho total dos dados para 4.769.

Em seguida, foram excluídos da base as ações que estão classificadas como financeiras, segundo a classificação do *Industry Classification Benchmark* (ICB). O ICB é um sistema de classificação das atividades industriais que foi criado pelo *Dow Jones* e pelo grupo *FTSE* em 2005 e é, atualmente, detido pelo *FTSE International*. O ICB utiliza um sistema de 10 indústrias, dividido em 19 super setores. Em seguida, foram retiradas as amostras que não apresentavam valor em alguma das variáveis selecionadas. Como base considerada final, temos 3.404 amostras com 28 variáveis.

Faz-se necessário pontuar algumas observações sobre os dados: a amostra está balanceada, pois há 1.734 resultados com a variável dependente igual a 1 (50,9%) e, algumas variáveis apresentam grande variância, sendo: ROIC em nível, valor da empresa sobre vendas em nível e rendimento do dividendo, demonstrando a presença de dados que se diferenciam drasticamente de todos os outros (*outliers*). Esses dados

podem causar anomalias nos resultados das classificações, uma vez que podem enviesar todo o resultado da análise. As variáveis com menor variância da amostra são: variação dos preços em 5 dias, 1 mês e 2 meses e o beta.

Para aplicarmos o ACP, as variáveis foram normalizadas uma vez que a técnica é sensível à escala das variáveis. Para tratarmos os *outliers*, foi aplicado a técnica de “winsorização” das variáveis, que consiste em aparar os valores extremos, substituindo-os pelos valores menores e maiores remanescentes na distribuição. Neste estudo, foram considerados extremos de 10% das observações de cada variável, sendo: 5% na parte inferior e 5% na parte superior de cada uma.

5. RESULTADOS

5.1. CLASSIFICADORES

Seguindo a metodologia descrita na seção 3 e a partir dos dados pré-processados, seguem os resultados passo a passo.

Tabela 7: Teste KMO e de Bartlett

Medida KMO	0,73
Teste de esfericidade de Bartlett	
p-valor	0,00
chi-quadrado	39.144
graus de liberdade	378

Fonte: elaboração própria

O teste de Bartlett demonstra que podemos rejeitar a hipótese nula (rejeição da matriz das correlações na população ser a identidade), uma vez que é observado um p-valor baixo para um nível de significância de 0,05. E no teste de KMO, o valor de 0,73 indica que a análise fatorial é apropriada, pois existe uma correlação média entre as variáveis.

Tabela 8: Análise de componentes principais

Componente principal	Autovalores	% de explicação da variância	% acumulado
1	129,24	17,52	17,52
2	104,08	11,37	28,89
3	93,27	9,13	38,02
4	78,29	6,43	44,45
5	71,47	5,36	49,81
6	66,60	4,66	54,47
7	66,28	4,61	59,08
8	62,74	4,13	63,21

9	58,04	3,53	66,74
10	56,52	3,35	70,09
11	55,51	3,23	73,32
12	53,15	2,96	76,28
13	50,81	2,70	78,98
14	49,35	2,56	81,54
15	47,27	2,35	83,89
16	45,98	2,22	86,11
17	45,24	2,15	88,26
18	41,84	1,84	90,10
19	40,67	1,74	91,84
20	38,95	1,60	93,44
21	35,32	1,31	94,75
22	33,48	1,18	95,93
23	29,35	0,90	96,83
24	28,74	0,87	97,70
25	26,25	0,72	98,42
26	25,30	0,67	99,09
27	22,67	0,53	99,62

Fonte: elaboração própria

Aplicando a técnica do ACP, é esperado que o menor número de componentes principais explique a maior parte da variância da amostra inicial. Analisando a tabela 8, percebemos que existe pouca concentração de variância. Não há uma regra para o número de componentes escolhidos e nem para o percentual acumulado de variância. Como exemplo, no trabalho de Hargraves e Mani (2015) são selecionados os 2 primeiros componentes, que explicam 94%. Com o intuito final de reduzir a dimensionalidade, iremos considerar os primeiros 18 componentes que explicam 90% da variância.

Aplicando os classificadores na amostra de teste, são obtidos os seguintes resultados:

Tabela 9: Resultado classificadores via matriz confusão

	SVM Linear	SVM Kernel RBF	Rede neural - <i>MLP</i>					
			1 camada oculta			2 camadas ocultas		
			10 neurônios	20 neurônios	37 neurônios	10 neurônios	20 neurônios	37 neurônios
Acurácia	54,8%	49,0%	50,8%	49,6%	48,5%	48,9%	48,2%	48,2%
Precisão	52,8%	48,1%	49,6%	48,3%	47,1%	47,7%	46,9%	47,2%
Sensitividade	69,0%	56,9%	51,8%	47,6%	45,8%	50,3%	47,3%	52,4%
Especificidade	58,3%	50,3%	52,1%	50,8%	49,7%	50,2%	49,4%	49,4%
F1-score	59,8%	52,1%	50,7%	48,0%	46,4%	49,0%	47,1%	49,6%

Fonte: elaboração própria

Segundo a tabela 9, o classificador que apresentou a maior acurácia foi o SVM com kernel linear, com 54,8% na amostra de teste; enquanto os demais classificadores apresentaram valores ao redor de 50%. Nota-se que para todos os algoritmos, a métrica sensibilidade (proporção de acertos nas observações positivas) apresentou valores altos, indicando a presença de mais falsos positivos do que falsos negativos. Em outras palavras, foram classificadas mais ações que tenderiam a desempenhar melhor que o mercado, quando essas não desempenharam.

5.2. CARACTERÍSTICAS DA AMOSTRA DE TESTE E DAS PREDIÇÕES

A amostra de teste possui 681 registros e o classificador rotulou 434 ações com predição 1 (retorno acima que o índice de referência) e 247 com predição 0.

Figura 7: Matriz confusão do SVM linear

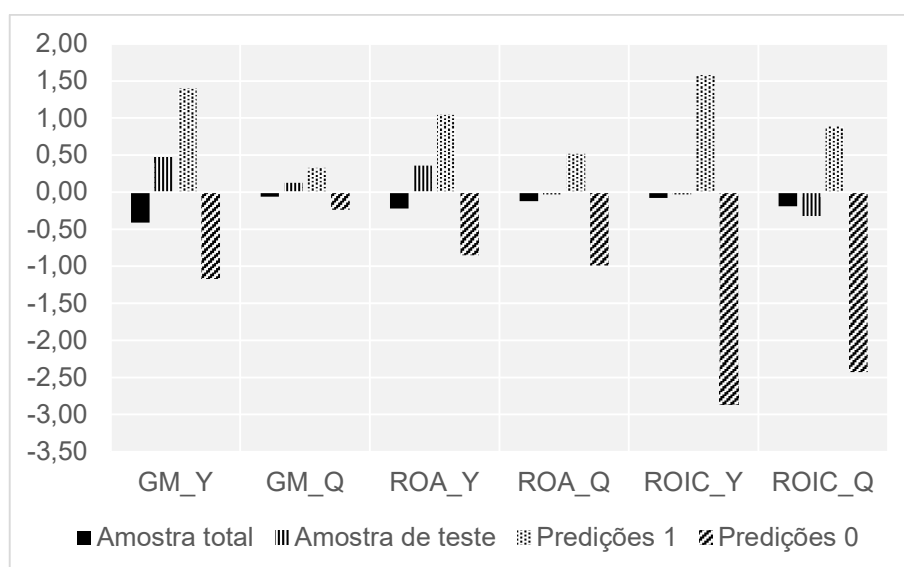
		Valores previstos	
		0	1
Valores reais	0	144	205
	1	103	229

Fonte: elaboração própria.

Na figura 7 é possível verificar que 373 amostras foram previstas corretamente, enquanto o classificador apresentou 205 falsos positivos e 103 falsos negativos.

Com o intuito de identificar os padrões utilizados pelo classificador, a seguir são verificadas as variáveis independentes que apresentaram as maiores diferenças em 4 bases distintas: i) a total, ii) a utilizada para teste, iii) apenas com previsões iguais a 1 e iv) apenas com previsões iguais 0. As figuras 8 até 11 representam os valores médios de cada variável em sua respectiva base. A tabela completa com todas as variáveis encontra-se no APÊNDICE A.

Figura 8: Comparação das variáveis de rentabilidade

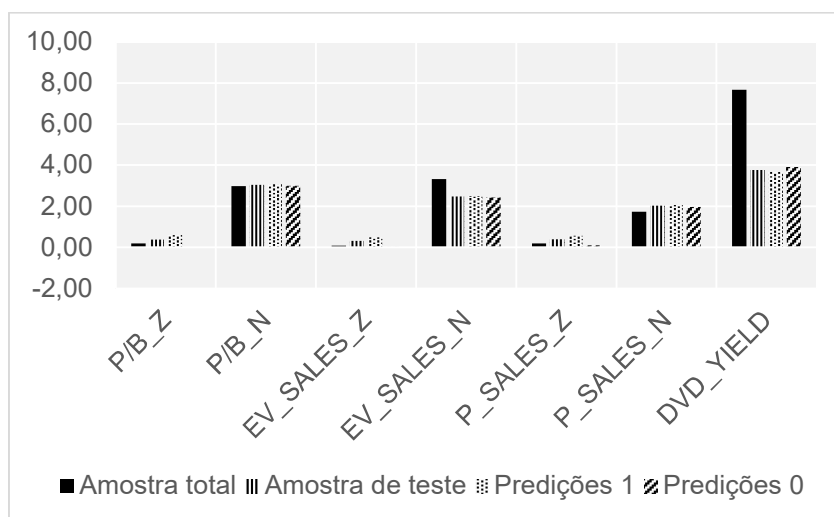


Fonte: elaboração própria.

É notório que as variáveis de rentabilidade apresentaram valores bem diferentes nas amostras de previsões, especificamente nas variáveis de evolução trimestral e anual da margem bruta, do ROIC e do retorno sobre o ativo. Enquanto na amostra predição 1 as variáveis apresentaram números positivos, ou seja, ganho de margem e aumento de rentabilidade, as variáveis da amostra predição 0 apresentaram decaimento. Essa

diferença sinaliza que a evolução das métricas de rentabilidade foram consideradas relevantes para a classificação.

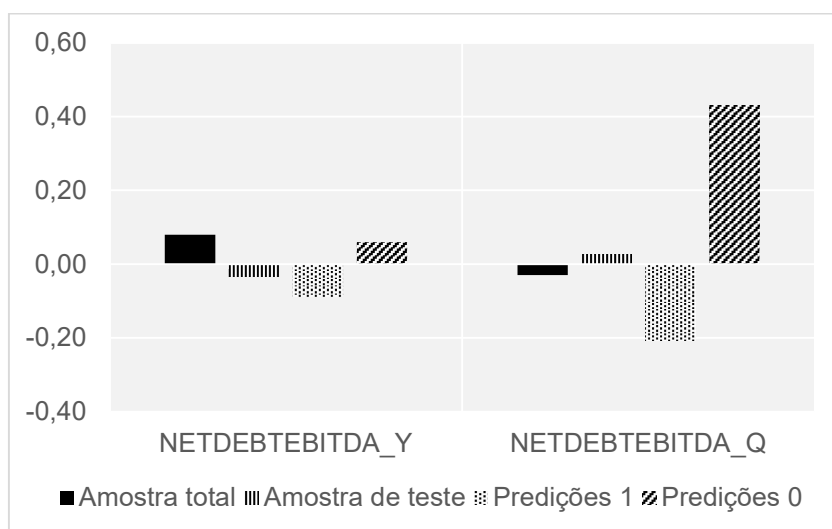
Figura 9: Comparação das variáveis de mensuração de valor



Fonte: elaboração própria.

Já as variáveis de mensuração de valor apresentaram diferenças mais discretas. Vale ressaltar que a amostra da predição 1 concentram ações com múltiplos de mensuração de valor mais elevados quando comparado a predição 0. Logo, o algoritmo SVM linear não estabeleceu um critério claro de padrão para essas variáveis.

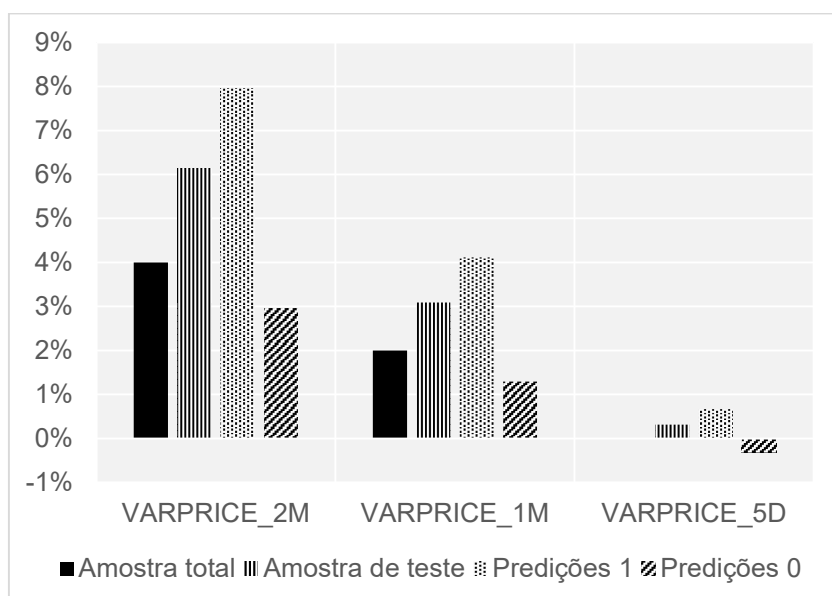
Figura 10: Comparação das variáveis de liquidez e endividamento



Fonte: elaboração própria.

Com relação as variáveis de liquidez e endividamento, as principais diferenças ocorreram nas de endividamento. Verifica-se que na predição 1 concentraram as empresas que apresentam redução em seus múltiplos, com relação ao trimestre anterior, e predição 0 com aumento desse múltiplo.

Figura 11: Comparação das variáveis de desempenho de mercado nas diferentes amostras



Fonte: elaboração própria.

Já nas variáveis de desempenho de mercado, o classificador priorizou as ações que apresentaram maiores retornos nos períodos de 2 meses, 1 mês e 5 dias com o preditor 1. Verificando assim, o padrão capturado no algoritmo.

Figura 12: Desempenho contra o índice de referência nas predições da amostra de teste

		Valores previstos	
		0	1
Valores reais	0	-10,48%	-8,82%
	1	11,19%	11,96%

Fonte: elaboração própria.

A figura 12 demonstra o desempenho do classificador com relação ao objetivo proposto deste trabalho. Os dados na tabela representam os valores médios da diferença entre o retorno da ação e o índice de referência, também chamado de excesso de retorno ou *alpha*, em suas respectivas classificações. Das amostras classificadas como 1 e, que nas observações reais pertenciam ao grupo 1 (com desempenho melhor que o IBrX 100), apresentaram um *alpha* médio de 11,96%. As ações classificadas como 0 (que apresentariam retorno abaixo do índice de referência) e que nas observações reais, também classificadas como 0, tiveram um *alpha* negativo 10,48%. No entanto, vale destacar as predições erradas, ou seja, os falsos positivos e falsos negativos. Os falsos positivos apresentaram *alpha* negativo de 8,82% e os falsos negativos exibiram *alpha* positivo de 11,19%.

5.3. BACK-TESTING

O período da amostra de teste se estendeu de 09/05/2017 até 31/10/2019, sendo o mesmo utilizado para o procedimento de *back-testing*. As ações definidas com predição 1 irão compor a Carteira 1, seguindo os critérios já explicitados na seção 3. **Metodologia**, subseção 3.6. **Back-testing e avaliação de desempenho da carteira**. As tabelas e figuras a seguir exibem as características e os resultados dessa carteira contra o índice de referência.

Tabela 10: Adições, exclusões e *turnover* da Carteira 1

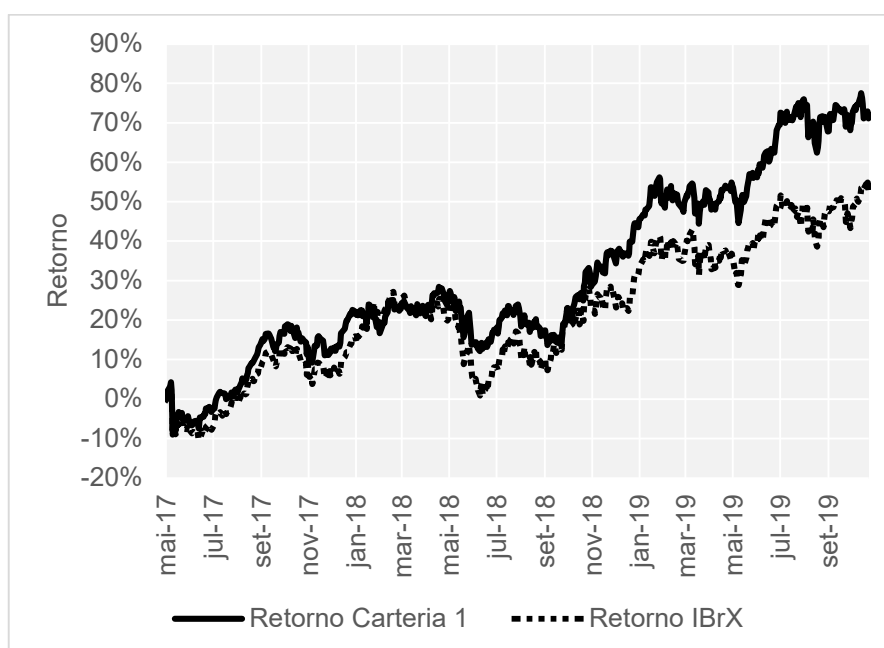
Mês	Adições	Exclusões	Quantidade de ações no final do período	<i>Turnover</i>
mai-17	12	0	12	2,5x
jun-17	1	0	13	0,1x
jul-17	16	-12	17	2,1x
ago-17	22	0	39	0,7x
set-17	0	-2	37	0,1x
out-17	17	-37	17	3,5x
nov-17	26	0	43	0,8x
dez-17	0	0	43	0,0x
jan-18	1	-42	2	2,7x
fev-18	21	-1	22	2,4x
mar-18	27	0	49	0,6x
abr-18	14	-7	56	0,4x
mai-18	35	-35	56	0,8x
jun-18	1	-7	50	0,2x
jul-18	16	-49	17	3,0x
ago-18	31	-1	47	1,0x
set-18	0	-1	46	0,0x
out-18	10	-46	10	2,8x
nov-18	32	0	42	1,0x
dez-18	0	0	42	0,0x
jan-19	0	-42	0	3,0x
fev-19	24	0	24	3,7x
mar-19	20	0	44	0,6x
abr-19	15	-3	56	0,3x
mai-19	38	-34	60	0,7x
jun-19	1	-13	48	0,2x

jul-19	9	-47	10	3,2x
ago-19	45	0	55	1,4x
set-19	0	-1	54	0,0x
out-19	0	-54	0	3,3x
Média	14	-14	34	1,4x

Fonte: elaboração própria.

Na tabela 10 é observado que a Carteira 1 apresentou troca de posições relevantes, em curto espaço de tempo, e carregou uma grande quantidade de ativos em determinados períodos. A Carteira 1 apresentou um volume negociado, dividido pelo patrimônio, também chamado de *turnover*, médio de 1,4x ao mês. Essa característica pode ser justificada devido a regra de vender o ativo adquirido em 50 dias úteis após a sua compra. É passível de verificação também, a grande quantidade de ações que a Carteira 1 possuía em alguns meses como abril/2018 e maio/2019. Tal característica reflete a ausência de filtros de exclusão para a seleção das ações.

Figura 13: Desempenho acumulado da Carteira 1 e do índice de IBrX 100



Fonte: elaboração própria.

Sobre a figura 13, a Carteira 1 exibiu retorno acumulado positivo de 71,13% enquanto o índice IBrX 100 obteve variação acumulada positiva de 53,32% no período analisado. Ou seja, a seleção de ações obteve um excesso de retorno sobre o índice de referência, ou *alpha*, de 17,80% positivo. Além disso, grande parte do *alpha* gerado se concentrou no ano de 2019.

Tabela 11: Análise de desempenho da Carteira 1 e do índice IBrX 100

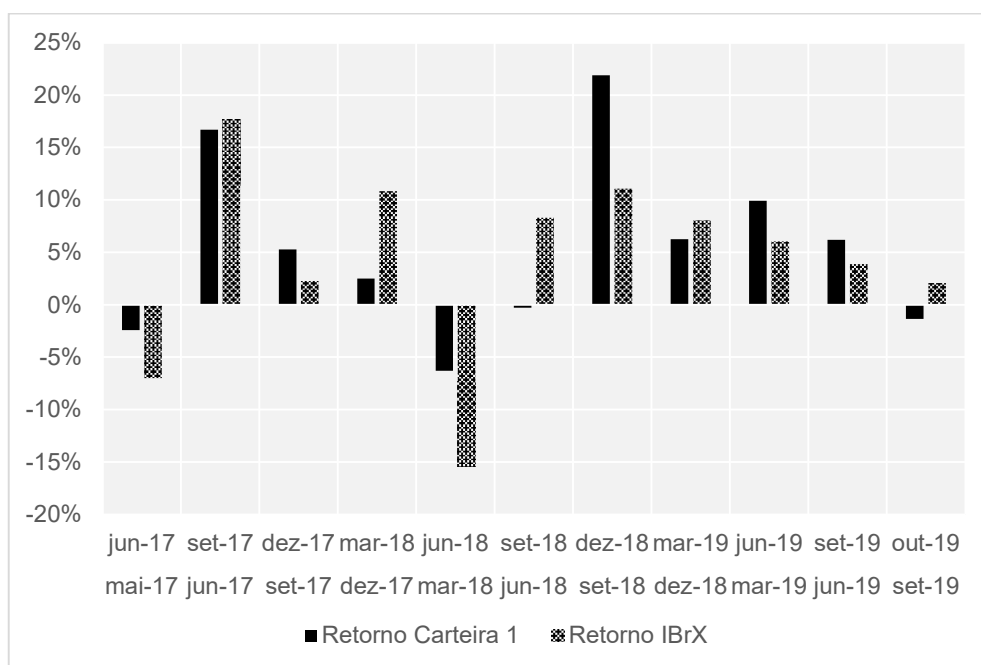
		Carteira 1		IBrX 100	
		Valor	Data		Data
Diário	Maior retorno absoluto	3,39%	19/05/2017	4,41%	08/10/2018
	Menor retorno absoluto	-10,01%	18/05/2017	-9,21%	18/05/2020
	Maior excesso de retorno (<i>alpha</i>)	2,62%	18/10/2018	---	---
	Menor excesso de retorno (<i>alpha</i>)	-2,35%	25/10/2019	---	---
	Ocorrências de <i>alphas</i> positivos	312	---	---	---
	Ocorrências de <i>alphas</i> negativos	290	---	---	---
Mensal	Maior retorno absoluto	12,82%	Outubro/2018	10,58%	Janeiro/2018
	Menor retorno absoluto	-7,26%	Maio/2018	-11,19%	Maio/2018
	Maior excesso de retorno (<i>alpha</i>)	4,31%	Julho/2019	---	---
	Menor excesso de retorno (<i>alpha</i>)	-9,88%	Janeiro/2018	---	---
	Ocorrências de <i>alphas</i> positivos	19	---	---	---
	Ocorrências de <i>alphas</i> negativos	11	---	---	---
Em todo o período	Maior retorno acumulado	77,54%	22/10/2019	54,91%	30/10/2019
	Menor retorno acumulado	-8,07%	22/05/2017	-10,01%	21/06/2017
	Perda Máxima *	-12,72%	---	-19,51%	---
Legenda: *Perda máxima: maior queda de valor do investimento, entre um pico e uma data posterior.					

Fonte: elaboração própria

A tabela 11 exibe dados sobre o desempenho da Carteira 1 e o índice de referência no período em destaque. Analisando os valores diários, percebe-se que o IBrX 100 apresentou o maior retorno absoluto, sendo 4,41% contra 3,39% da Carteira 1; porém o mesmo índice apresentou um menor retorno absoluto de -9,21%, contra -10,01% da Carteira 1. Com relação aos dados de excessos de retornos, ocorreram mais dias com *alphas* positivos do que *alphas* negativos, 312 e 290, respectivamente. Com relação

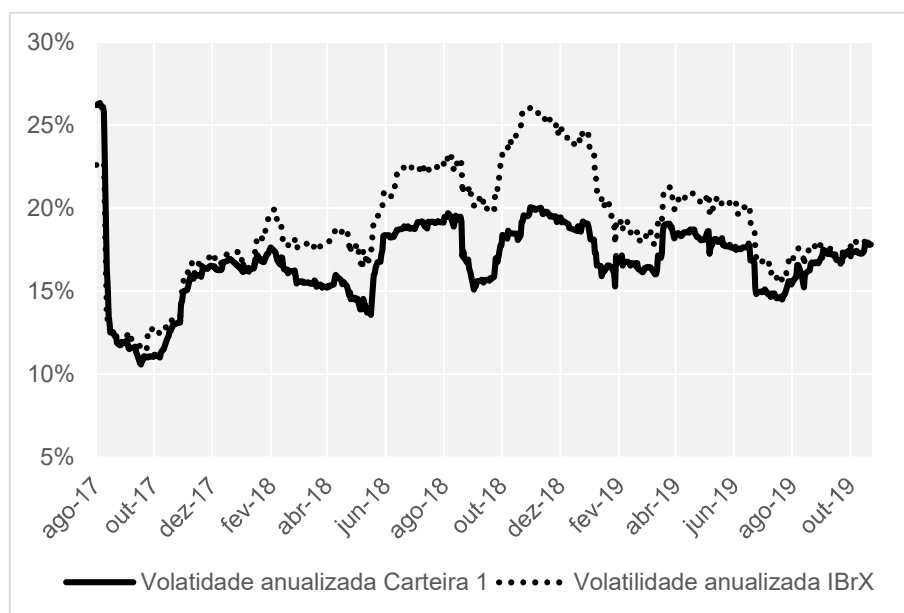
aos dados mensais, a Carteira 1 apresentou números superiores de retornos absolutos comparado ao índice de referência. Porém, vale destacar o menor excesso de retorno da Carteira 1 de -9,88% em janeiro de 2018, ou seja, neste mês o desempenho da carteira foi quase 10% inferior em relação ao índice. No período acumulado, as 3 métricas analisadas demonstram números superiores para a Carteira 1 quando comparado ao índice.

Figura 14: Desempenho trimestral da Carteira 1 e do índice IBrX 100



Na figura 14 é notório que o retorno da Carteira 1 foi superior ao IBrX 100 em 6 trimestres: 2º trimestre de 2017, 4º trimestre de 2017, 2º trimestre de 2018, 4º trimestre de 2018, 2º e 3º trimestres de 2019. O período que apresentou o maior excesso de retorno foi o 4º trimestre de 2018, o mesmo período que definiu as eleições presidenciais no Brasil.

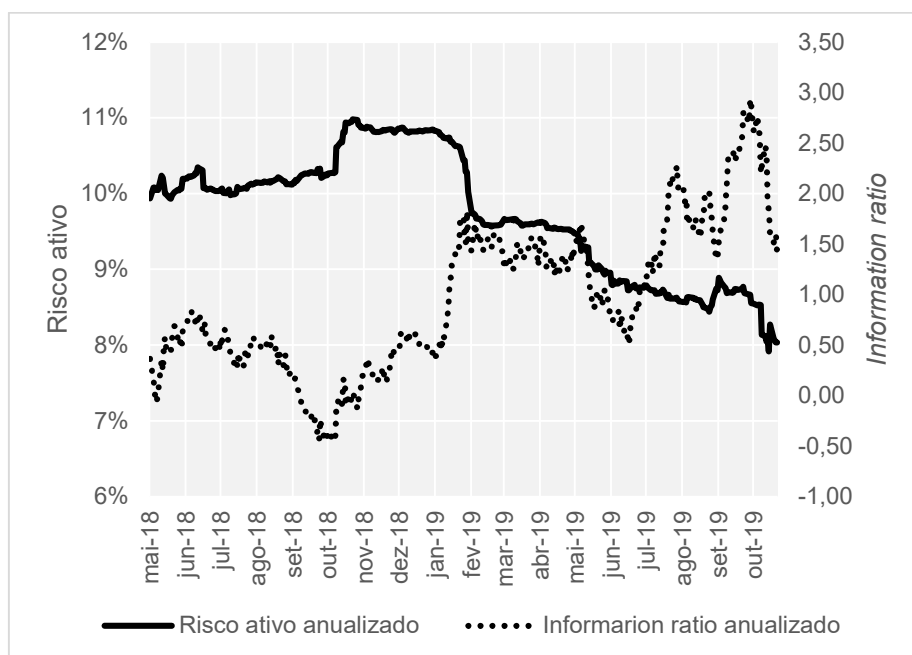
Figura 15: Evolução da volatilidade da Carteira 1 e do IBrX 100



Fonte: elaboração própria.

A figura 15 apresenta a evolução da volatilidade da Carteira 1 e do índice IBrX 100 anualizada. Observa-se que na maior parte do período analisado, a Carteira 1 apresentou uma volatilidade menor. O desempenho do classificador e o fato da Carteira 1 possuir pesos iguais para todos os ativos em cada período deve ter contribuído para tal comportamento.

Figura 16: Evolução do risco ativo e do *information ratio* da Carteira 1



Fonte: elaboração própria.

O risco ativo da Carteira 1 permaneceu no intervalo entre 11% e 8%, considerado alto para uma carteira que apresenta um número elevado de ações em determinados períodos. Já o *information ratio* foi mais volátil: ficou no intervalo de -0,50 a +1,00 no ano de 2018, de +0,50 a +3,00 em 2019.

6. CONCLUSÕES

O objetivo deste trabalho é propor um modelo de seleção de ações que superem o índice IBrX 100, com base em seus registros contábeis e indicadores de mercado, a partir da utilização das técnicas de *support vector machine* e redes neurais - identificadas ao longo deste estudo como classificadores.

Os dados foram coletados e tratados, chegando a um total de 28 variáveis com mais de 3 mil amostras. Os testes de Kaiser-Meyer-Olkin e de esfericidade de Bartlett apresentaram resultados satisfatórios a fim de verificar a consistência para análise fatorial.

A técnica de ACP foi eficaz em reduzir o total de variáveis de 29 para 18 principais componentes na base completa. Ao analisarmos os resultados dos classificadores na amostra de teste, foi verificado que a técnica de *SVM* com kernel linear apresentou a maior acurácia, com 54,8%. No entanto, o mesmo classificador apresentou sensibilidade de 69%, indicando a presença de mais falsos positivos do que falsos negativos. Em outras palavras, foram classificadas mais ações que tenderiam a desempenhar melhor que o mercado, quando essas não desempenharam.

Ao analisarmos o desempenho do classificador *SVM* linear na amostra de teste, as ações com predição 1 concentraram empresas que apresentaram aumento de rentabilidade, redução de endividamento, e que exibiram retornos positivos nos períodos de 2 meses, 1 mês e 5 dias. Já as variáveis de mensuração de valor apresentaram diferenças mais discretas nas predições 0 e 1. Ou seja, o algoritmo *SVM* linear não estabeleceu um critério claro de classificação para essas variáveis.

Os resultados do *back-testing* demonstraram que a Carteira 1 apresentou retorno positivo de 71,13%, enquanto o índice IBrX 100 retornou + 53,32% no período referente a amostra de teste. Os níveis de volatilidade da Carteira 1 se mantiveram menor que o IBrX 100 e, o desempenho da carteira com relação ao seu risco ativo foi relevante no período de 2019.

Uma possível crítica do resultado ocorre no período considerado para o *back-testing* - o índice de referência e a maior parte das ações elegíveis do classificador apresentaram retornos positivos. Também é observado a presença de mais falsos positivos do que falsos negativos na amostra de teste, demonstrando o viés de resultados dos classificadores, apesar das amostras de treinamentos estarem balanceadas.

Como sugestão para trabalhos futuros, faz sentido incluir variáveis que tragam noções de expectativa de mercado, como lucro líquido esperado. Pode-se esperar que mudanças nas projeções de lucros de determinada empresa possa influenciar no comportamento do preço da ação. Outrossim, a utilização de outros métodos de aprendizado que podem vir a ser melhores classificadores para este conjunto de dados. Também se devem se considerar outros períodos para a realização do procedimento de *back-testing*, que apresentem tendências positivos e negativos a fim de verificar o desempenho dos ativos selecionados em todos os ciclos.

REFERÊNCIAS

ALBERG, J.; LIPTON, Z. C. *Improving Factor-Based quantitative Investing by Forecasting Company Fundamentals*. 31st Conference on Neural Information Processing Systems (NIPS), 2018.

ALMEIDA, H.; CAMPELLO, M.; LARANJEIRA, B.; WEISBENNER, S. *Corporate Debt Maturity and the Real Effects of the 2007 Credit Crisis*. Working Paper No. 14990, NATIONAL BUREAU OF ECONOMIC RESEARCH, 2009.

ARNOTT, R.; HARVEY, C. R.; MARKOWITZ, H. *A Backtesting Protocol in the Era of Machine Learning*. The Journal of Financial Data Science, pp. 64-74, 2019.

BELLMAN, R. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

BURGES, C. J. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery 2, pp. 121-167, 1998.

CAO, L. J.; TAY, E. H. *Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting*. IEEE Transactions on Neural Networks, Vol 14, No6, pp. 1506-1518, 2003.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification* (2 ed.). Wiley-Interscience, 2000.

FAN, A.; PALANISWAMI, M. *Stock Selection using Support Vector Machines*. University of Melbourne, Department of EEE, 2001.

GRAHAM, B. *The Intelligent Investor*. Harper & Brothers, 1949.

GRANTHAM, R. J. *The Case for Quality - The Danger of Junk*. White Paper, GMO, 2004.

GREENBLATT, J. *The Little Book That Still Beats the Markets*. Hoboken, New Jersey: John Wiley & Sons, 2010.

HARGRAVES, C. A.; MANI, C. K. *The Selection of Winning Stocks Using Principal Component Analysis*. American Institute of Science. American Journal of Marketing Research, 2015.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation* (2 ed.). Pearson Prentice Hall, 1999.

HERBRICH, R. *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press, 2002.

JOLLIFFE, I. T. *Principal Component Analysis* (2 ed.). Springer-Verlag New York, 2002.

KAISER, H. F. *The Application of Eletronic Computers to Factor Analysis*. University of Illinois. Educational and Psychological Measurement, Vol. 20, No. 1, 1960.

KITANI, E. C. *Mapeamento e Visualização de Dados em Alta Dimensão com Mapas Auto-Organizados*. Tese de Doutorado, Universidade de São Paulo, Departamento de Engenharia de Sistemas Eletrônicos, São Paulo, 2013.

LORENA, A. C.; CARVALHO, A. C. *Uma Introdução às Support Vector Machines*. RITA, Vol XIV, N.2, 2007.

MACIEL, L. S.; BALLINI, R. *Neural Networks Applied to Stock Market Forecasting: an Empirical Analysis*. UNICAMP, Economics Institute (IE). Journal of the Brazilian Neural Networks Society, Vol.8, Iss. 1, pp. 3-22, 2010.

MBELEDUGU, N. N.; ODOH, M.; UMEH, M. N. *Stock Feature Extraction Using Principal Component Analysis*. Nnamdi Azikiwe University, Department of Computer Science. International Conference on Computer Technology and Science. doi:10.7763/ IPCSIT.2012.V47.44, 2012.

NOVY-MARX, R. *The Other Side of Value: The Gross Profitability Premium*. Journal of Financial Economics 108 (1), pp. 1-28, 2013.

NOVY-MARX, R. *Quality Investing*. Unpublished Results, 2014.

PIOTROSKI, J. D. *Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers*. Journal of Accounting Research, pp. 1-41, 2000.

PRADO, M. L. *The 10 Reasons Most Machine Learning Funds Fail*. The Journal of Portfolio Management, 44 (6) 120-133. doi:<https://doi.org/10.3905/jpm.2018.44.6.120>, 2018.

QI, M.; MADDALA, G. S. *Economic Factors and the Stock Market: a New Perspective*. Journal of Forecasting. p. 151-166, 1999.

SHARPE, W. F. *Mutual Fund Performance*. The Journal of Business, pp.119-138, 1966.

SHEELA, K. G.; DEEPA, S. N. *Review on Methods to Fix Number of Hidden Neurons in Neural Networks*. Anna University. Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 425740, 11 pages. doi:10.1155/2013/425740, 2013.

SLOAN, R. G. *Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?* The Accounting Review, Vol, 71, No. 3, pp. 289-315.

TREYNOR, J. L., & BLACK, F. (1973). *How to Use Security Analysis to Improve Portfolio Selection*. The Journal of Business, Vol.46 No. 1, pp. 66-86. The University of Chicago Press, 1996.

YANG, L.; REA, W.; REA, A. *Stock Selection with Principal Component Analysis*. Journal of Investment Strategies 5: 1-21, 2016.

YU, H.; CHEN, R.; ZHANG, G. *A SVM Stock Selection Model within PCA*. *Procedia Computer Science* 31 - 406 - 412. Elsevier B.V. doi:10.1016/j.procs.2014.05.284, 2014.

APÊNDICE

APÊNDICE A - Comparação das variáveis nas amostras: total, de teste, de predição 1 e de predição 0.

Variável	Amostra total			Amostra de teste			Predições 1			Predições 0		
	Média	Mediana	Desvio padrão	Média	Mediana	Desvio padrão	Média	Mediana	Desvio padrão	Média	Mediana	Desvio padrão
<i>GM_Y</i>	-0,41	-0,22	8,14	0,47	0,04	9,77	1,40	0,61	7,00	-1,17	-1,13	13,17
<i>GM_Q</i>	-0,06	-0,02	7,63	0,12	0,02	7,35	0,33	0,29	5,78	-0,24	-0,46	9,51
<i>ROA_Y</i>	-0,22	-0,16	4,77	0,36	0,39	5,11	1,04	0,73	4,50	-0,85	-0,30	5,85
<i>ROA_Q</i>	-0,12	-0,05	1,92	-0,03	0,05	2,29	0,52	0,33	1,52	-0,99	-0,57	3,00
<i>ROA_N</i>	7,37	6,45	6,86	6,67	5,26	6,51	6,98	5,50	5,92	6,11	4,67	7,41
<i>ROIC_Y</i>	-0,08	-0,09	8,35	-0,03	0,65	8,28	1,58	1,21	6,68	-2,87	-1,20	9,91
<i>ROIC_Q</i>	-0,19	-0,05	4,12	-0,32	0,04	4,86	0,89	0,48	2,44	-2,43	-1,01	6,91
<i>ROIC_N</i>	16,85	13,10	19,41	15,51	11,51	22,25	14,76	11,98	14,94	16,81	10,35	31,19
<i>P/B_Z</i>	0,18	0,22	1,14	0,36	0,50	1,08	0,59	0,79	1,02	-0,03	-0,08	1,07
<i>P/B_N</i>	2,98	1,68	3,98	3,04	1,92	3,09	3,06	2,07	2,86	2,99	1,72	3,46
<i>EV_SALES_Z</i>	0,08	0,05	1,15	0,31	0,48	1,11	0,48	0,66	1,08	0,01	0,03	1,12
<i>EV_SALES_N</i>	3,32	1,91	22,58	2,46	2,01	1,64	2,49	2,06	1,57	2,42	1,92	1,74
<i>P_SALES_Z</i>	0,18	0,27	1,18	0,39	0,57	1,10	0,56	0,71	1,07	0,09	0,08	1,10
<i>P_SALES_N</i>	1,72	1,23	2,03	2,02	1,59	1,56	2,05	1,61	1,54	1,96	1,52	1,60
<i>DVD_YIELD</i>	7,67	3,71	39,68	3,72	2,84	3,31	3,62	2,61	3,27	3,89	3,06	3,40
<i>EBITDA_Y</i>	0,40	1,00	0,92	0,46	1,00	0,89	0,67	1,00	0,74	0,08	1,00	1,00
<i>EBITDA_Q</i>	0,32	1,00	0,95	0,38	1,00	0,92	0,95	1,00	0,31	-0,61	-1,00	0,79
<i>EPS_Y</i>	0,13	1,00	0,99	0,16	1,00	0,99	0,59	1,00	0,81	-0,59	-1,00	0,81
<i>EPS_Q</i>	0,04	1,00	1,00	0,03	1,00	1,00	0,12	1,00	0,99	-0,12	-1,00	1,00
<i>NETDEBTEBITDA_Y</i>	0,08	0,01	7,09	-0,03	-0,05	4,41	-0,09	-0,06	3,83	0,06	-0,04	5,28
<i>NETDEBTEBITDA_Q</i>	-0,03	0,01	6,26	0,02	0,00	3,38	-0,21	-0,02	2,26	0,43	0,07	4,73
<i>NETDEBTEBITDA_N</i>	1,25	1,31	5,22	1,45	1,53	3,50	1,38	1,47	2,24	1,57	1,73	5,01
<i>CURRATIO_Y</i>	0,01	0,00	0,74	-0,01	0,01	0,72	-0,03	-0,01	0,66	0,01	0,03	0,81
<i>CURRATIO_Q</i>	0,01	0,01	0,49	0,01	0,00	0,65	0,03	0,00	0,56	-0,03	0,01	0,79
<i>VARPRICE_2M</i>	0,04	0,03	0,17	0,06	0,05	0,14	0,08	0,06	0,14	0,03	0,03	0,15
<i>VARPRICE_1M</i>	0,02	0,02	0,11	0,03	0,02	0,10	0,04	0,03	0,09	0,01	0,00	0,11
<i>VARPRICE_5D</i>	0,00	0,00	0,05	0,00	0,00	0,05	0,01	0,01	0,05	0,00	0,00	0,05
<i>RAW_BETA</i>	0,70	0,67	0,30	0,74	0,71	0,32	0,70	0,68	0,31	0,81	0,79	0,33