**INSTITUTO FEDERAL DE EDUCAÇÃO CIÊNCIA E TECNOLOGIA DE MINAS GERAIS – CAMPUS FORMIGA**

**MESTRADO PROFISSIONAL EM ADMINISTRAÇÃO**

**NATAN FELIPE SILVA**

# ANÁLISE DE DESEMPENHO DE PORTFÓLIOS QUE COMBINAM MODELOS DE *MACHINE LEARNING* E OTIMIZAÇÃO MULTIOBJETIVO NO MERCADO DE AÇÕES BRASILEIRO

**FORMIGA – MG**

**2023**

Natan Felipe Silva

# ANÁLISE DE DESEMPENHO DE PORTFÓLIOS QUE COMBINAM MODELOS DE *MACHINE LEARNING* E OTIMIZAÇÃO MULTIOBJETIVO NO MERCADO DE AÇÕES BRASILEIRO

Trabalho de conclusão de curso apresentado ao Programa de Pós-graduação em Administração do Instituto Federal de Educação, Ciências e Tecnologia de Minas Gerais – IFMG - Campus Formiga, como requisito para conclusão do curso de Mestrado Profissional em Administração.

Linha de pesquisa: Finanças corporativas e investimentos.

Orientador: Prof. Dr. Lélis Pedro de Andrade

Coorientador: Prof. Dr. Washington Santos Silva

FORMIGA – MG

2023

Natan Felipe Silva

ANÁLISE DE DESEMPENHO DE PORTFÓLIOS QUE COMBINAM MODELOS DE *MACHINE LEARNING* E OTIMIZAÇÃO MULTIOBJETIVO NO MERCADO DE AÇÕES BRASILEIRO

Trabalho de conclusão de curso apresentado ao Programa de Pós-graduação em Administração do Instituto Federal de Educação, Ciências e Tecnologia de Minas Gerais – IFMG - Campus Formiga, como requisito para conclusão do curso de Mestrado Profissional em Administração.

Aprovado em _____/_____/_____ pela banca examinadora:

BANCA EXAMINADORA

---

Prof. Dr. Lélis Pedro de Andrade (orientador)

---

Prof. Dr. Washington Santos Silva (coorientador)

---

Prof. Dr. Paulo Henrique Sales Guimarães (Avaliador externo – UFLA)

---

Prof. Dra. Maísa Kely de Melo (Avaliadora interna – IFMG Campus Formiga)

*S586a* Silva, Natan Felipe.

Análise de desempenho de portfólios que combinam modelos de Machine Learning e otimização multiobjetivo no mercado de ações brasileiro / Natan Felipe Silva. - Formiga, 2024

84 p. : il. color.

Dissertação (Mestrado Profissional em Administração) – Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Formiga, 2024.

Orientador: Dr. Lélis Pedro de Andrade.

1.Investimentos. 2. Seleção de carteiras de investimentos. 3. Otimização de portfólio. 4. Markowitz. 5. Machine learning. I. Silva, Natan Felipe. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Formiga. III. Título.

CDD:332.024

**Catalogação: Lívia Renata Santos- CRB/6-2561**

“As pessoas que vencem nesse mundo são as pessoas que procuram as circunstâncias que precisam e, quando não as encontram, as criam.” George Bernard Shaw

Dedico este trabalho a todos que, de alguma forma, se esforçaram para contribuir para uma sociedade melhor e com menos desigualdades sociais, e que sejam capazes de olhar para o futuro com esperança e otimismo.

## AGRADECIMENTOS

Agradeço a Deus pelo apoio e incentivo, assim como pela Sua presença constante. A Nossa Senhora, agradeço pelas inúmeras intercessões junto ao seu Filho, possibilitando que este trabalho fosse concluído de maneira honrosa.

Expresso minha profunda gratidão aos meus queridos pais, Ourivaldo e Maria Lucia, e às minhas amadas irmãs, Ana Paula e Carla, assim como aos meus adoráveis sobrinhos, Arthur, Augusto e Alice. Essa família sempre me apoiou incondicionalmente, não medindo esforços nem orações para o meu sucesso. À minha namorada, Mariele, agradeço pela paciência e apoio constante ao longo desta jornada.

Aos professores que contribuíram para a minha formação durante o mestrado, meu sincero agradecimento. Em especial, expresso minha gratidão aos meus orientadores, o Prof. Dr. Lélis, e ao coorientador, Prof. Dr. Washington, que foram fundamentais ao longo desta caminhada. Aos membros da banca avaliadora, agradeço por aceitarem o convite e por estarem dispostos a contribuir com esta dissertação. Agradeço a todos os amigos e profissionais que, por meio de incentivos e experiências compartilhadas, contribuíram para enriquecer e aprimorar este trabalho.

## *RESUMO*

A construção de um portfólio de investimento através da seleção criteriosa de ações é uma atividade crucial para os investidores. Nesse contexto, os métodos de inteligência artificial despontam como ferramentas fundamentais para apoiar as decisões dos investidores. O objetivo principal desta pesquisa é comparar diferentes métodos de *machine learning* para realizar a pré-seleção de ativos e analisar a combinação desses métodos na avaliação do desempenho de carteiras otimizadas. Essas carteiras serão integradas a um modelo de otimização multiobjetivo que busca maximizar o retorno e minimizar o risco. Para atingir essa meta, a dissertação desdobrou-se nos seguintes objetivos específicos, cada um estruturado como um artigo:i) Otimização de portfólio de investimento: uma revisão bibliométrica (Produto Bibliográfico 1); ii) Seleção de ativos e otimização de portfólios de investimentos com métodos de inteligência artificial: uma revisão sistemática e bibliométrica da literatura (Produto Bibliográfico 2); iii) Análise de desempenho de portfólios que combinam modelos de machine learning e otimização multiobjetivo no mercado de ações brasileiro (Produto Bibliográfico 3). O artigo empírico proposto nesta dissertação foi fundamentado nos dois artigos de revisão bibliométrica. Utilizando os modelos de *machine learning*, nomeadamente: a) *Random Forest*, b) *Multilayer Perceptron* (MLP), e c) *Extreme Gradient Boosting* (XGBoost), cada modelo foi treinado e validado para realizar a pré-seleção de ações com base em indicadores financeiros. Posteriormente, após a seleção de ações feita por cada modelo, foi conduzida a otimização do portfólio para determinar o percentual de alocação em cada ativo. É importante destacar que as carteiras selecionadas pelos modelos de inteligência artificial foram testadas por meio de *backtesting*, no qual foram calculados os seguintes indicadores de desempenho: i) índice de Sharpe, ii) índice de Treynor, iii) Alfa de Jensen e iv) VAR. O artigo empírico evidenciou que a combinação de modelos de *machine learning* e um modelo multiobjetivo gerou indicadores de desempenho superiores em comparação com os portfólios testados sem o uso combinado da técnica de machine learning e otimização multiobjetivo. Finalmente, como produto técnico, apresenta-se um algoritmo utilizado no artigo empírico (Produto Técnico/Tecnológico 1).



**Palavras-chaves**: investimentos; seleção de carteiras de investimentos; otimização de portfólio; Markowitz; machine learning.

**Abstract**

Building an investment portfolio through careful stock selection is a crucial activity for investors. In this context, artificial intelligence methods emerge as fundamental tools to support investors' decisions. The main aim of this research is to compare different machine learning methods to perform asset pre-selection and to analyze the combination of these methods in evaluating the performance of optimized portfolios. These portfolios are integrated into a multi-objective optimization model that seeks to maximize return and minimize risk. To achieve this goal, the dissertation was divided into the following specific objectives, each structured as an article:i) Investment portfolio optimization: a bibliometric review (Bibliographic Product 1); ii) Asset selection and optimization of investment portfolios with artificial intelligence methods: a systematic and bibliometric review of the literature (Bibliographic Product 2); iii) Performance analysis of portfolios that combine machine learning models and multi-objective optimization in the Brazilian stock market (Bibliographic Product 3). The empirical article proposed in this dissertation was based on the two bibliometric review articles. Using machine learning models, namely: a) Random Forest, b) Multilayer Perceptron (MLP), and c) Extreme Gradient Boosting (XGBoost), each model was trained and validated to pre-select stocks based on financial indicators. Subsequently, after the selection of shares made by each model, portfolio optimization was conducted to determine the allocation percentage in each asset. It is important to highlight that the portfolios selected by the artificial intelligence models were tested through backtesting, in which the following performance indicators were calculated: i) Sharpe index, ii) Treynor index, iii) Jensen's Alpha and iv) VAR . The empirical article showed that the combination of machine learning models and a multi-objective model can generate results significantly superior to the market benchmark. Finally, as a technical product, the dissertation provides the script and database used in the empirical article (Technological Product 1).

**Keywords:** investments; selection of investment portfolios; portfolio optimization, Markowitz; machine learning.

# SUMÁRIO

# 1.INTRODUÇÃO

O problema de alocação ótima de portfólio de investimentos foi estudado inicialmente por Markowitz (1952), que propôs a teoria moderna de portfólio, a qual busca a alocação ótima dos pesos em cada ativo com base na análise do retorno esperado, variância e covariância dos ativos. O estudo de Markowitz (1952) foi o percursor para o desenvolvimento de diversas pesquisas no campo de otimização e seleção de portfólio de investimentos, visto que os investidores buscam equilibrar a relação risco e retorno, diminuindo as incertezas de mercado, sendo que fatores políticos e macroeconômicos influenciam no desempenho do portfólio (Deng et al., 2022; Wang et al., 2020).

Considerando a hipótese de eficiência de mercado, em que todos os investidores possuem as mesmas informações e o mercado se movimenta por um *random walk* (passeio aleatório) (Fama, 1998, 1991), a busca por um portfólio com desempenho acima do retorno esperado seria inviável. Sendo assim, considera-se que os estudos de otimização e seleção de portfólio, além de solucionar um problema de investimento maximizando o retorno e minimizando o risco, também buscam oferecer resultados acerta do teste de eficiência de mercado.

Alguns estudos buscando ampliar a metodologia proposta por Markowitz (1952) têm aplicado métodos para realizar a pré-seleção de ativos antes de realizar a otimização do portfólio, a exemplo dos modelos de *machine learning* (aprendizado de máquina), *deep learning* (aprendizado profundo) e de métodos de séries temporais, que foram aplicados para selecionar ativos na bolsa da China e posteriormente foram combinados com modelo média – variância, buscando qual modelo traria melhor resultado (Ma et al., 2021).O Extreme Gradient Boosting (XGBoost) também foi aplicado na pré-seleção de ações com base em indicadores financeiros para uma posterior otimização por meio do modelo média – variância (CHEN et al., 2021). Da mesma forma, o algoritmo *random forecast* (floresta aleatória) foi utilizado na pré-seleção de ações (Ballings et al., 2015). A regressão SVM (Support Vector Machine) também foi aplicada ao mercado de ações (Matías and Reboredo, 2012).

A seleção de investimento via um comitê SVM para posterior otimização de portfólio foi utilizada por (Paiva et al., 2019).O algoritmo de redes neurais LSTM, em conjunto com simulação de Monte Carlo na geração de portfolios aleatórios, foi aplicado à seleção e otimização de portfólio de ações A rede neural de memória curta LSTM foi aplicada para selecionar ações para investimentos do S&P 500 e chegou-se à conclusão que modelos de otimização com pré-seleção superam modelos de otimização de portfólio com a seleção livre (Fischer and Krauss, 2018).

Buscando contribuir com o avanço da literatura, esta pesquisa propõe estudar a otimização de portfolio com métodos de pré-seleção usando *machine learning* na bolsa de valores brasileira. Forão utilizados os modelos i) *Random Forest*, ii) *Multilayer Perceptron (MLP)*, iii) *Extreme Gradient Boosting (XGBoost)*, sendo que os ativos selecionados por cada modelo tiveram a otimização realizada por um modelo de otimização de portfolio que viza maximizar o retorno e minimizar o risco. Os dados das carterias otimizadas foram testados com dados fora da amostra, de modo que serão calculados os indicadores de desempenho das carteiras.

A estrutura da dissertação está dividida em quatro partes, além da seção introdutória que ainda apresenta os objetivos do estudo. A seção dois trará uma revisão bibliométrica acerca de seleção e otimização de portfólio com o intuito de levantar um panorama geral acerca do tema. A seção três apresenta um estudo bibliométrico focado no uso da inteligência artificial e na seleção e otimização de portfólio de investimentos. A terceira seção irá demostrar o artigo empírico. A quarta seção apresenta o produto técnico/ tecnológico ouriundo do artigo empirico.

## 2. CONSIDERAÇÕES FINAIS

O objetivo da dissertação é avaliar o desempenho de portfólios compostos por ativos pré-selecionados por meio de modelos de machine learning e, posteriormente, otimizados pelo modelo multiobjetivo. Para tanto, foram realizadas duas revisões bibliométricas: uma no âmbito da otimização de portfólio de investimento e outra considerando a integração da otimização de portfólio de investimento com a inteligência artificial. Em ambas as revisões, observou-se um crescente interesse na integração dos métodos de inteligência artificial na área de otimização de portfólio de investimentos.

As duas revisões bibliométricas foram conduzidas para fornecer embasamento a um trabalho empírico que combinou a otimização multiobjetivo de portfólios de investimentos com métodos de machine learning. Nesse contexto, foram aplicados os métodos de machine learning: Multilayer Perception, XGBoost e Random Forest, os quais foram combinados com um modelo multiobjetivo que buscava maximizar o retorno e minimizar o risco simultaneamente. Como dados de entrada, uma série de indicadores financeiros foi calculada, escolhidos após uma extensa revisão de literatura e posteriormente submetidos a um processo de feature selection. Cada modelo foi treinado e validado por meio de um método de validação cruzada. Posteriormente, cada modelo selecionou um grupo de ativos, os quais foram integrados ao modelo de otimização multiobjetivo.

A partir da implementação do modelo de otimização, foram calculadas diversas métricas para avaliar o desempenho dos portfólios. Em todas essas métricas, os portfólios otimizados superaram o benchmark de mercado. Destacam-se, em particular, o retorno anual e o índice de Sharpe, nos quais os portfólios apresentaram um desempenho superior. Além disso, métricas como o índice de Treynor, o beta e o *Alfa de Jensen* foram avaliadas para os portfólios, todas indicando maior eficiência em relação ao mercado. A performance dos portfólios também foi comparada ao longo do tempo com o benchmark, e os portfólios otimizados consistentemente superaram a carteira de mercado.

A dissertação buscou contribuir para a área de seleção e otimização de portfólios de investimentos, tanto por meio dos artigos de revisão bibliométrica e sistemática quanto com o artigo empírico, cuja metodologia é replicável a outros mercados. Os artigos de revisão bibliométrica já foram publicados e apresentados em congresso, e a expectativa é submetê-los para publicação em um periódico em breve. A partir do estudo empírico, surgirão dois trabalhos adicionais: um artigo empírico a ser submetido a um periódico internacional e um produto técnico composto por um algoritmo, a ser submetido ao Code Ocean. Essa submissão vinculará a publicação ao artigo, tornando a metodologia replicável.

Apesar de o estudo atingir o objetivo proposto, ressaltam-se algumas melhorias que podem ser implementadas como sugestão de trabalhos futuros : i) testar outros modelos de *machine learning* para realizar a pré-seleção de ações; ii) testar outros algoritmos para resolver o problema de otimização, como NSGA II (BARROSO; CARDOSO; MELO, 2021; DE MELO; CARDOSO; JESUS, 2022; PIMENTA et al., 2018); iii) aplicar restrição de cardinalidade aos portfólios, como em (BARROSO; CARDOSO; MELO, 2021; PAIVA et al., 2019); iv) aplicar a metodologia proposta no presente estudo a outros mercados.

## 2.1 CONTRIBUIÇÕES

A execução desta pesquisa de mestrado produziu as contribuições para a área de seleção e otimização de portfólios de investimentos descritas na Tabela 1:

Tabela 1- Contribuições do estudo

| **Título do produto** | **Autores** | **Tipo de produção** |
|---|---|---|
| Otimização de portfólio de investimento: uma revisão bibliométrica | (SILVA et al., 2022) | Publicado Convibra 2022 |
| Seleção de ativos e otimização de portfólios de investimentos com métodos de inteligência artificial:uma revisão sistemática e bibliométrica da literatura | (SILVA et al.; 2023) | Publicado Semead 2023 |

Fonte: Elaborado pelos autores.

Além das publicações mencionadas na Tabela 1, a expectativa é que os artigos de revisão bibliométrica sejam submetidos em periódicos, uma vez que passaram por um processo de condensação após a revisão pelos pareceristas nos congressos. Também se espera que o Produto 3, que não está listado na Tabela 1, mas é apresentado na dissertação, seja submetido a um periódico internacional, acompanhado do algoritimo (script) que será submetido à plataforma Code Ocean gerando-se assim um produto técnico.

## REFERÊNCIAS


BALLINGS, M. et al. Evaluating multiple classifiers for stock price direction prediction. **Expert Systems with Applications**, v. 42, n. 20, p. 7046–7056, 2015.

CHEN, W. et al. Mean–variance portfolio optimization using machine learning-based stock price prediction. **Applied Soft Computing**, v. 100, p. 106943, mar. 2021.

DENG, X. et al. Non-dominated sorting genetic algorithm-II for possibilistic mean-semiabsolute deviation-Yager entropy portfolio model with complex real-world constraints. **Mathematics and Computers in Simulation**, v. 202, p. 59–78, dez. 2022.

FAMA, E. F. Efficient Capital Markets: II. **The Journal of Finance**, v. 46, n. 5, 1991.

FAMA, E. F. Market efficiency, long-term returns, and behavioral finance. **Journal of Financial Economics**, v. 49, n. 3, 1998.

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. **European Journal of Operational Research**, v. 270, n. 2, p. 654–669, 2018.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, 2021.

MARKOWITZ, H. Portfolio selection. **The Journal of Finance**, v. 7, n. 1, p. 77–91, 1952.

MATÍAS, J. M.; REBOREDO, J. C. Forecasting performance of nonlinear models for intraday stock returns. **Journal of Forecasting**, v. 31, n. 2, p. 172–188, 2012.

PAIVA, F. D. et al. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. **Expert Systems with Applications**, v. 115, p. 635–655, 2019.

SILVA, N. F. et al. **Otimização de portfólio de investimento: uma revisão bibliométrica**. XIX CONGRESSO VIRTUAL DE ADMINISTRAÇÃO , dez. 2022.

SILVA, N. F. et al. An integrated CRITIC and Grey Relational Analysis approach for investment portfolio selection. **Decision Analytics Journal**, p. 100285, jul. 2023.

TA, V.-D.; LIU, C.-M.; TADESSE, D. A. Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. **Applied Sciences (Switzerland)**, v. 10, n. 2, 2020.

WANG, W. et al. Portfolio formation with preselection using deep learning from long-term financial data. **Expert Systems with Applications**, v. 143, p. 113042, abr. 2020.

## PRODUTO 1 (Bibliográfico): Otimização e seleção de portfólio de investimentos: uma revisão bibliométrica e sistemática

**Resumo:** Os métodos de otimização e seleção de portfólios de investimento têm sido estudados ao longo dos anos desde que Markowitz propôs o modelo média – variância, em 1952. Mesmo diante de reconhecida repercussão do trabalho, inúmeros têm sido os trabalhos que buscam aprimorar o modelo original ou a criação de novas metodologias. O objetivo desses estudos é buscar soluções cada vez mais satisfatórias em um tempo cada vez menor sem, contudo, chegarem a uma definição a respeito do método mais adequado para se definir um portfólio ótimo. Sob esse escopo, origina-se o presente trabalho, que tem como objetivo realizar uma análise bibliométrica da literatura na área de otimização de portfólio de investimentos com intuito de identificar os principais estudos sobre o tema, bem como as lacunas e tendências de pesquisas. Para tanto, a pesquisa foi realizada na base *Scopus,* contemplando todo o período de dados disponível. Posteriormente, os trabalhos foram analisados e a bibliometria realizada, utilizando o pacote Bibliometrix da linguagem R. A pesquisa evidenciou um total de 3.149 artigos a respeito do tema. Por meio de uma análise do número de publicações ao longo dos anos, observou-se que houve um crescimento de 22% da média de publicações dos últimos 5 anos em relação à média geral de publicações. Também foram realizadas análises em relação aos principais autores, países e fontes mais relevantes, que identificaram uma predominância de publicações de alto impacto dos pesquisadores especialmente da China e Estados Unidos. Por fim, foi feita uma análise acerca dos métodos aplicados que convergem para um crescimento na aplicação de Lógica Fuzzy, *machine learning, deep learning* na temática de otimização de portfólio.

**Palavras-chave:** Otimização de portfólio de investimentos; revisão bibliométrica; modelo média – variância.

**Title: Investment portfolio optimization and selection: a bibliometric and systematic review**

**Abstract:** Investment portfolio optimization methods have been studied over the years since Markowitz proposed the mean-variance model in 1952. Despite the recognized repercussion of the work, there have been countless works that seek to improve the original model or the creation of new methodologies aiming at increasingly satisfactory solutions in an increasingly shorter time without, however, arriving at a definition regarding the most adequate method to define an optimal portfolio. Under this scope originates the present work that aims to carry out a bibliometric analysis of the literature in the area of investment portfolio optimization in order to identify the main studies on the subject, as well as gaps and research trends. For this purpose, the survey was carried out in the Scopus database covering the entire period of data available. Subsequently, the works were analyzed and bibliometrics performed using the R language Bibliometrix package. The research

showed a total of 3149 articles on the subject. Through an analysis of the number of publications over the years, it was observed that there was a 22% growth in the average of publications in the last 5 years in relation to the general average of publications. Analyzes were also carried out in relation to the main authors, countries and most relevant sources, which identified a predominance of high impact publications by researchers, especially from China and the United States. Finally, an analysis was made of the applied methods that converge-to a growth in the application of Fuzzy Logic, machine learning, deep learning in the theme of portfolio optimization.

**Keyword**: Investment portfolio optimization; bibliometric review; mean – variance model.

## 1. Introdução

Os métodos quantitativos aplicados a finanças estão atraindo a atenção dos pesquisadores devido à possibilidade de solução de problemas, seja na otimização de portfólio, na previsão de retorno, ou mesmo na gestão de risco (Corberán-Vallet et al., 2023). Os métodos de otimização de portfólios vêm sendo estudados ao longo dos anos, visto que os investidores buscam constantemente maior eficiência entre risco e retorno em suas decisões de investimento (Kolm et al., 2014).

Um dos primeiros métodos de otimização de portfólios foi desenvolvido ainda na década de 1950. Na ocasião, Markowitz (1952) propôs a teoria moderna de portfólio que, por meio da diversificação dos ativos na composição de uma carteira, teve o objetivo de minimizar o risco de uma carteira. O método de inserções dessas variáveis nas carteiras ficou conhecido como modelo de média-variância. Desde então, diversos estudos (Kalayci et al., 2019) têm aplicado o modelo de média variância na composição de carteiras de investimentos. No entanto, estudos (H.R. Golmakani and Fazel, 2011; Mehlawat, 2016) têm utilizado novos métodos na busca de otimizar o portfólio, sendo que alguns (X. Cui et al., 2012; Utz et al., 2014) combinam o modelo de média-variância (Markowitz, 1952) com outros métodos quantitativos, oriundos da estatística, pesquisa operacional, séries temporais e computação científica, cujo intuito é o de obter desempenhos considerados melhores na relação risco e retorno.

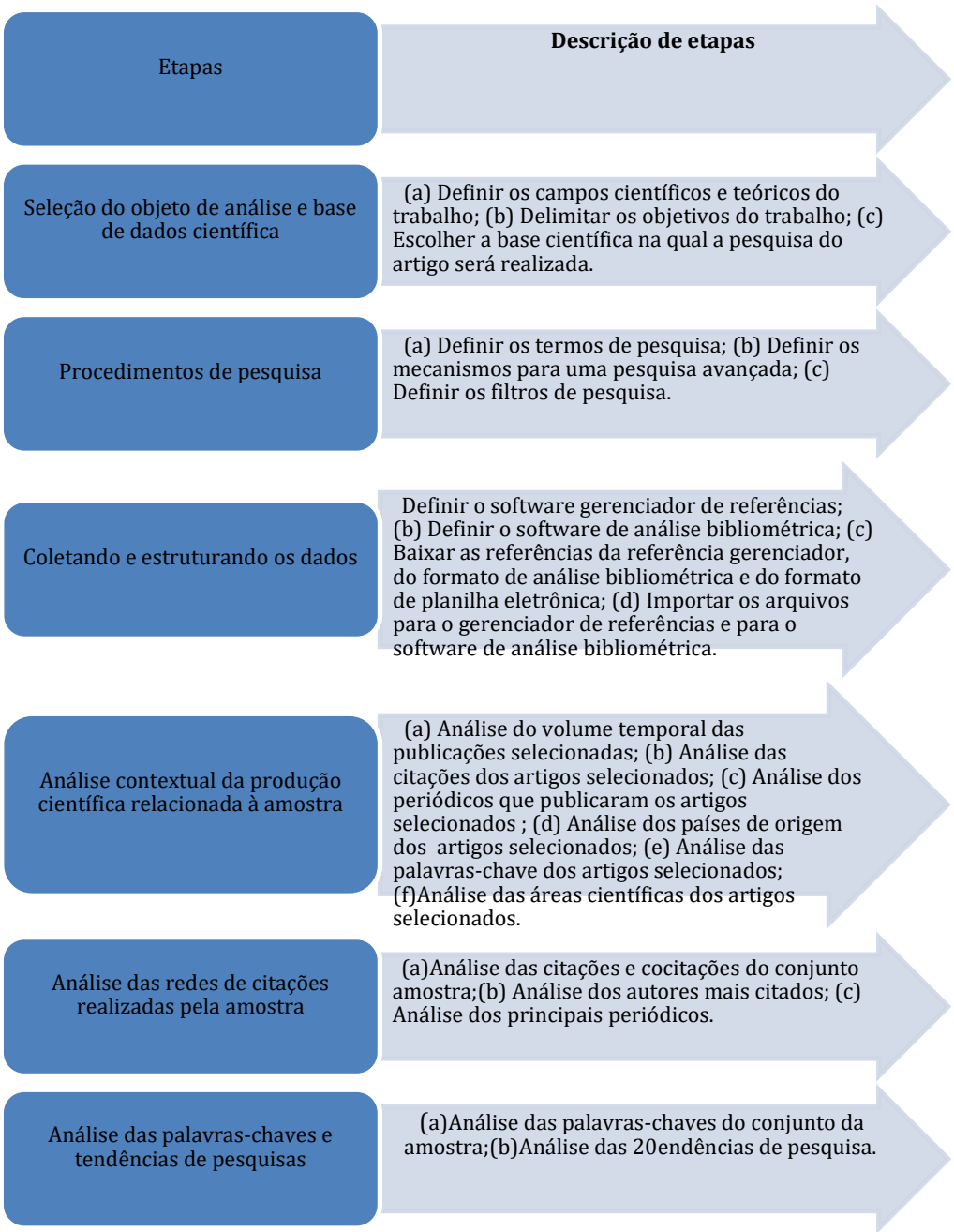
Por outro lado, certas carteiras, mesmo que constituídas com base em critérios científicos, apresentam desempenhos abaixo do esperado. Assim, mesmo com o crescimento das tecnologias e da aplicação de métodos de análise de dados, se faz necessário compreender como eles estão sendo aplicados na composição de carteiras de investimentos e, principalmente, qual é a tendência de pesquisa nessa área. Com isso, o presente estudo busca responder a seguinte questão: Quais são os principais métodos e as tendências de pesquisas na área de otimização de portfólio de investimentos?

Diante desse contexto, o trabalho emerge com o objetivo de realizar uma análise bibliométrica na área de otimização de portfólio de investimentos, com a finalidade de identificar os principais estudos sobre o tema, bem como as lacunas e tendências de pesquisas na área.

## 2. Metodologia

A pesquisa caracteriza-se como uma análise bibliométrica, que foi realizada utilizando os artigos extraídos da base *Scopus*, sendo também classificada como descritiva e dotada de abordagem quantitativa (Costa et al., 2017). Para realizar a análise bibliométrica, foram seguidas as etapas listadas na Figura 1 (COSTA et al., 2017).

Figura 1- Etapas da Montagem da Pesquisa e Análise Bibliométrica.




Fonte: Elaborado pelos autores - adaptado a partir de Costa et al. (2017).

Observam-se na Figura 1 as etapas que serão seguidas no decorrer da pesquisa. A primeira etapa de delimitar o objetivo e o campo da pesquisa ocorreu na introdução do trabalho. A base de dados que será utilizada para realizar o estudo será a base *Scopus*. A metodologia aplicada nas etapas subsequentes será descrita nos próximos tópicos.

## 2.1. Procedimentos Metodológicos

Para definição e identificação dos termos de busca e palavras-chaves considerando a temática de otimização de portfólio de investimentos, foram utilizadas expressões booleanas com intuito de trazer à tona o maior número de trabalhos que contemplam a temática.

A busca foi realizada no título, resumo e palavras-chave seguindo a seguinte linha de comando:

(Portfolio AND ((selection OR management OR optimization) AND (assessment OR metrics OR measures OR estimation) AND (investment OR finance)).

Na terceira etapa da Figura 1, coleta e estruturação de dados para análise bibliométrica foi utilizado o pacote Bibliometrix do software R, programa utilizado na literatura conforme (Paltrinieri et al., 2019; Pattnaik et al., 2020) pela praticidade de quantificar e processar os dados. Para realizar as análises utilizando o Bibliometix, seguiu-se o estudo de Aria; Cuccurullo (2017), que apresenta uma série de possibilidades de análises bibliométricas com o pacote e suas respectivas funções.

A quarta etapa da Figura 1 consiste em realizar uma análise contextual da produção científica selecionada pela amostra cujo objetivo é conhecer as publicações selecionadas pela busca. Para tanto, primeiramente foi realizada uma análise sobre o número de artigos da amostra, reunidos ao longo dos anos. Além disso, também foram analisadas as citações dos artigos selecionados, a fim de identificar as publicações mais relevantes da amostra. Ademais, foram realizadas análises sobre os periódicos que publicaram os artigos selecionados, focando também seus países de origem.

A quinta etapa da análise bibliométrica, Figura 1, consiste em analisar e discutir as redes de citações realizadas pela amostra. Essa análise será realizada com o objetivo de conhecer os estudos e autores mais importantes citados pelos trabalhos selecionados pelos critérios de busca. Para tanto, inicialmente foi realizada uma análise das obras mais citadas e, logo em seguida, realizada uma discussão, focada na relação entre os autores principais e a quantidade total de suas obras que passam a ser citadas pelos artigos da amostra.

A sexta etapa da análise bibliométrica, Figura 1, consiste em analisar e discutir as palavras chaves que aparecem com mais frequência no estudo, identificar tendências de pesquisas e buscar lacunas relacionadas aos principais métodos de otimização que estão sendo aplicados.

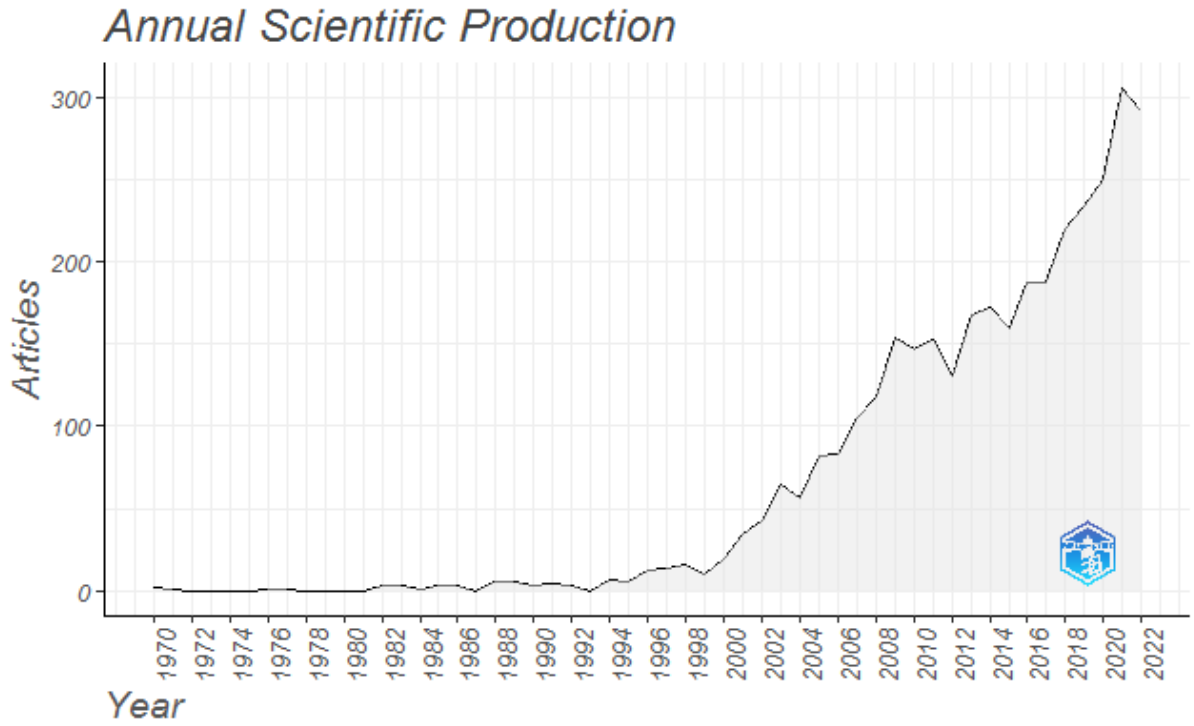
# 3 Resultados e Discussões

## 3.1 Análise contextual da amostra e da produção científica

A pesquisa foi realizada considerando o período de 1970 a 2022, todo o período disponível na base *Scopus*. Abusca principal retornou um total de 3.149 artigos. Com o objetivo de conhecer o

recorte temporal das publicações inicialmente identificadas, elaborou-se um gráfico a partir de um critério de busca previamente estabelecido, que pode ser observado na Figura 2, mostrando o número de artigos publicados ao longo do tempo.

Figura 2: Análise das publicações ordenadas por quantidade




Fonte: Elaborada pelos autores com os dados da pesquisa.

Observa-se na Figura 2 um crescimento no número de publicações, principalmente nos últimos 10 anos. Os dois primeiros trabalhos da amostra foram publicados em 1970, um intitulado "*The portfolio analysis of multiperiod capital investment under conditions of risk*",publicado na revista "*Engineering Economist"* com 4 citações na *Scopus*, e o outro "*Market allocation under uncertainty*", publicado na revista "*European Economic Review",* que possui 49 citações. Ambos os estudos discutem sobre otimização e seleção de portfólio de investimentos e sobre análise de risco (Dreze, 1970; Levy and Sarnat, 1970).

Dentre os artigos selecionados com base nos critérios de busca previamente estabelecidos, alguns se destacaram de acordo com o número de citações na base *Scopus*. Portanto, a Tabela 1 apresenta os estudos com mais citações conforme a busca realizada.

Tabela 1 - Cinco artigos mais citados

| Título | Autor | Citações |
|---|---|---|

| | | |
|---|---|---|
| Conditional value-at-risk for general loss distributions | Rockafellar e Uryasev (2002) | 2165 |
| A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms | DeMiguel et al. (2009) | 455 |
| Worst-case value-at-risk and robust portfolio optimization: a conic programming approach | Ghaoui, Oks e Oustry (2003) | 407 |
| High dimensional covariance matrix estimation using a factor model | Fana , Fanb , Lv (2008) | 312 |
| Stock return predictabilityand model uncertainty | Avramov (2002) | 268 |

Fonte: Elaborada pelos autores com os dados da pesquisa.

Com 2.165 citações, o artigo de Rockafellar e Uryasev (2002) apresenta o CVAR como medida de avaliação de risco.Através das análises, os autores chegaram à conclusão de que essa medida traz uma contribuição para a avaliação do risco de um portfólio de investimentos. O segundo artigo mais citado, com 455 citações na base *Scopus*, de Demiguel et al., (2009), fornece um *framework* para encontrar carteiras com alto desempenho por meio do modelo média-variância, mas com restrições adicionais, dentre elas, que o vetor peso da carteira seja menor que um determinado limite estipulado. O terceiro artigo mais citado na temática da base *Scopus* é dos autores (Ghaoui; Oks e Oustry, (2003), que propõem um modelo de otimização de carteira de investimentos baseado no Modelo Média- variância, propondo modificações no modelo com o intuito de minimizar o risco e tornar o modelo mais robusto. O quarto estudo, com 312 citações, é dos pesquisadores Fan e Fan; Lv, (2008), que estudam o impacto da estimativa da matriz de covariância utilizada no modelo média – variância na alocação ótima de portfólio. O quinto e último estudo listado na Tabela 1 é a pesquisa de Avramov, (2002), que propõe um modelo baseado em estatística Bayesiana para prever retornos das ações, o que acaba trazendo contribuição para área de otimização de portfólio de investimentos, pois uma das etapas mais importantes na construção de carteiras ótimas é estimar os retornos.

A pesquisa também analisou os periódicos com maior relevância no tema, sendo considerados os índices H, G e M, TC (total de citações e NP (número de publicações.) A Tabela 2 apresenta os periódicos com maior impacto levando em consideração esses parâmetros.

Tabela 2- 10 periódicos mais ranqueados conforme número de citações

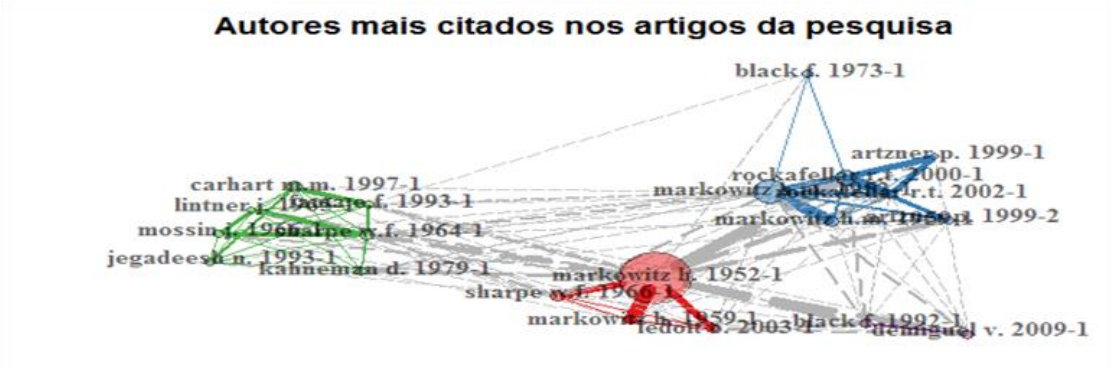
| Periódico | Índice H | Índice G | Índice M | TC | NP |
|---|---|---|---|---|---|
| EUROPEAN JOURNAL OF OPERATIONAL RESEARCH | 20 | 37 | 1,67 | 1422 | 49 |
| EXPERT SYSTEMS WITH APPLICATIONS | 21 | 31 | 1,75 | 1045 | 39 |
| ACM INTERNATIONAL CONFERENCE PROCEEDING SERIES | 2 | 4 | 0,29 | 22 | 20 |
| ANNALS OF OPERATIONS RESEARCH | 10 | 14 | 1,00 | 221 | 19 |
| LECTURE NOTES IN COMPUTER SCIENCE | 5 | 8 | 0,42 | 76 | 19 |
| JOURNAL OF INDUSTRIAL AND MANAGEMENT OPTIMIZATION | 4 | 5 | 0,36 | 39 | 17 |
| MATHEMATICAL PROBLEMS IN ENGINEERING | 5 | 8 | 0,42 | 73 | 17 |
| XITONG GONGCHENG LILUN YU SHIJIAN/SYSTEM ENGINEERING THEORY AND PRACTICE | 4 | 4 | 0,36 | 41 | 17 |
| JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS | 6 | 11 | 0,50 | 137 | 15 |
| QUANTITATIVE FINANCE | 6 | 12 | 0,50 | 160 | 15 |

Fonte: elaborada pelos autores com os dados da pesquisa.

### 3.2 Análise das redes da produção científica

A análise de redes mostrou os autores que foram mais citados e seus respectivos estudos. A Figura 3 apresenta os autores mais citados.

Figura 3: Análise das publicações ordenadas por quantidade




Fonte: Elaborada pelos autores com os dados da pesquisa.

Observa-se na Figura 3 que os dois trabalhos de Markowitz estão no centro de citações para embasar os estudos:o estudo basilar que gerou o modelo media-variância *"Portfolion selection"* Markowitz (1952) e o estudo *"Portfolio Selection: Efficient Diversification of Investments "*que trata da importância de diversificar uma carteira para minimizar o risco (MARKOWITZ, 1959). Outro trabalho basilar para fundamentar a otimização de carteira de investimentos é o de Sharpe (1966), que desenvolveu uma métrica para avaliar risco e retorno que ficou conhecida como índice de Sharpe.

Além de realizar a análise dos autores que foram mais citados nas pesquisas, é importante ressaltar os autores que estão publicando pesquisas mais relevantes e cujos estudos apareceram na amostra de dados. A Tabela 3 apresenta a análise de autoria conforme o índice H, G e M que são fornecidos pela base *Scopus*. Abaixo segue a Tabela 3.

Tabela 3- 10 autores mais bem ranqueados

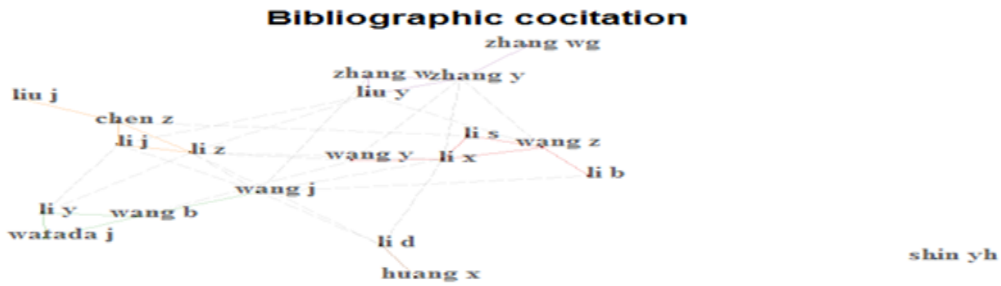
| Autor | Índice H | Índice G | Índice M | TC | NP |
|---|---|---|---|---|---|
| LI X | 12 | 20 | 1 | 426 | 27 |
| LI Y | 5 | 10 | 0,455 | 132 | 21 |
| SHIN YH | 6 | 8 | 0,5 | 90 | 20 |
| LI Z | 5 | 12 | 0,417 | 157 | 19 |
| NA NA | 1 | 1 | 0,091 | 4 | 19 |
| HUANG X | 7 | 14 | 0,583 | 212 | 18 |
| ZHANG Y | 6 | 10 | 0,545 | 119 | 18 |
| CHEN Z | 6 | 10 | 0,545 | 109 | 17 |

| | | | | | |
|---|---|---|---|---|---|
| LI D | 8 | 16 | 0,727 | 314 | 16 |

Fonte: elaborada pelos autores com os dados da pesquisa.

Observa-se que o autor LI X é o mais citado, com 426 citações, e possui o maior número de publicações, além de possuir também maiores índices M e G. O autor pertence ao departamento da *School of Economics and Management* da *Beijing University of Chemical Technology* da China. Dentre seus estudos, podemos destacar o *"Multi-period mean-semi-entropy portfolio management with transaction costs and bankruptcy control"*, que investiga o problema de otimização de portfólio de investimentos utilizando Lógica Fuzzy (Zhou et al., 2021). Outra pesquisa recente do autor é a pesquisa *"Equilibrium strategy for a multi-period weighted mean-variance portfolio selection in a Markov"*, que testa a combinação do modelo média – variância com cadeias de Markov (Ge et al., 2021).O estudo foi também publicado em parceria com o pesquisador Li Z, que também aparece na Tabela 3. Na Figura 4, podemos observar a relação de citação entre os autores.

Figura 4: Citação entre os autores




Fonte: Elaborado pelos autores com os dados da pesquisa.

Observa-se na Figura 4 que os autores têm citado uns aos outros. Percebe-se que os autores presentes na Tabela 3 estão correlacionados e que alguns grupos citam uns aos outros. Um ponto a ressaltar é que há uma predominância de autores chineses listados na Figura 4. A Tabela 4 ressalta essa predominância através da análise de citações por países.

Tabela 4- Total de citações por país

| **País** | **Total de citações** | **Média de Citação por artigo** |
|---|---|---|
| CHINA | 2953 | 7,17 |
| USA | 1546 | 9,91 |

| | | |
|---|---|---|
| SPAIN | 1353 | 22,18 |
| IRAN | 871 | 15,02 |
| TAIWAN | 714 | 11,16 |
| ITALY | 703 | 12,78 |
| GERMANY | 664 | 13,02 |
| INDIA | 625 | 7,81 |
| UNITED KINGDOM | 576 | 8,73 |
| HONG KONG | 565 | 14,49 |
| BRAZIL | 427 | 7,24 |
| KOREA | 290 | 5,69 |
| AUSTRALIA | 277 | 8,39 |
| FRANCE | 269 | 9,28 |
| NETHERLANDS | 241 | 24,10 |
| GREECE | 227 | 7,83 |
| CANADA | 221 | 7,13 |
| JAPAN | 220 | 6,88 |
| AUSTRIA | 203 | 18,46 |
| SAUDI ARABIA | 155 | 25,83 |

Fonte: Elaborada pelos autores com os dados da pesquisa.

Observa-se na Tabela 4 a predominância da China, sendo que possui quase o dobro de trabalhos citados em relação ao segundo país com maior número de citações (Estados Unidos).

### 2.1.3.3 Análise das palavras chaves e tendências de pesquisas

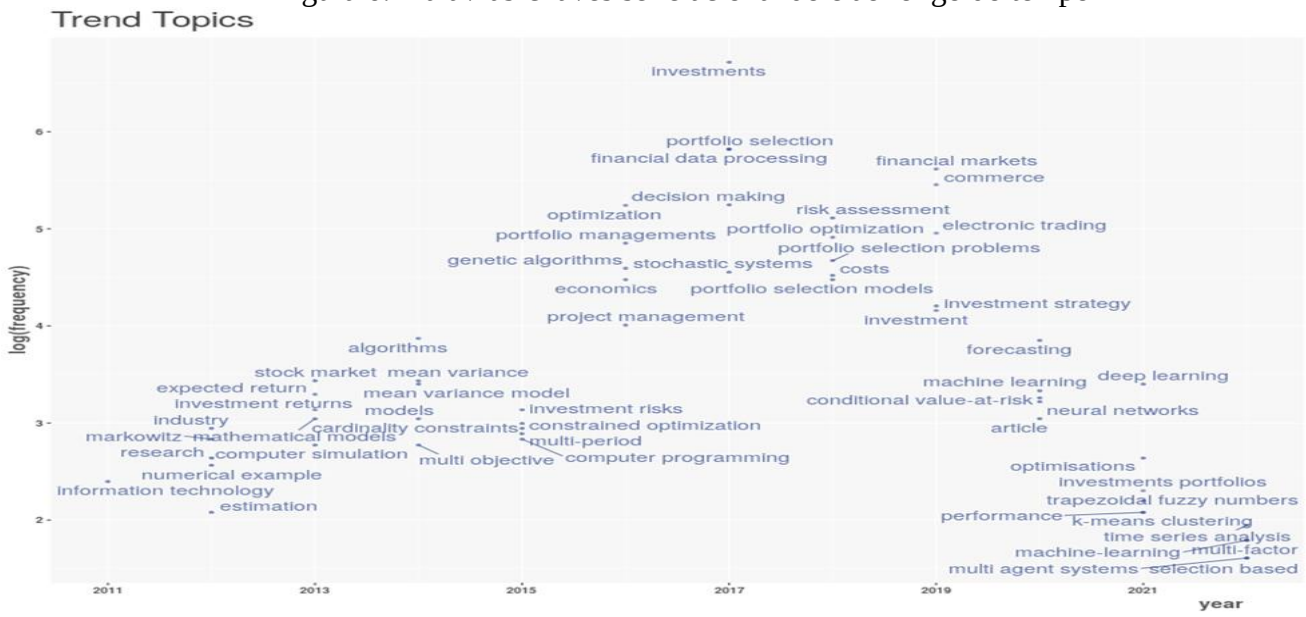
A Figura 5 apresenta neste tópico as *Keywords* (palavras-chaves) que apareceram com mais frequências na amostra do presente estudo, a partir da base *Scopus*. Em destaque, podemos notar os termos “seleção de portfólios”, “processamento de dados financeiros”, “decisões de mercado” e “risco”, dentre outras obtidas, conforme apresentado na Figura 5.

Figura 5: Nuvem de palavras-chaves


Fonte: Elaborado pelos autores com os dados da pesquisa.

Na Figura 5, observa-se a ocorrência principal de palavras-chaves como *"portfolio management*" e *"portfolio optmization*". Um ponto a se destacar é que esses termos estão ligados à chave de busca que foi determinada na seção 3.2. Outras palavras-chaves que apareceram com frequência e merecem destaque são: *"efficient frontier"* ou fronteira eficiente, que está ligada ao modelo de média – variância proposto por Markowitz, (1952), "Sharpe ratio" ou índice de Sharpe, que está ligado ao trabalho de (Sharpe, 1964) que também é citado na Figura 3. Outras palavras como "value risk" e "investimento" também estão ligadas diretamente com o tema e com o modelo média – variância. Também devemos destacar a presença de "*machine learning"* (aprendizado de máquina) e *"genetic algorithm*" (algoritimo genético). Ambas estão ligadas ao advento da inteligência artificial, que pode ser aplicada à otimização de portfólio de investimentos. Sendo assim, com o intuito de identificar tendências de pesquisa, foi realizada uma análise fatorial levando em conta a ocorrência das palavras-chaves ao longo do tempo. Tal análise pode ser observada na Figura 6.

Figura 6: Palavras-chaves se relacionando e ao longo do tempo




Fonte: Elaborado pelos autores com os dados da pesquisa.

A Figura 6 levou em consideração apenas os 10 últimos anos com intuito de dar ênfase às pesquisas mais recentes. Observa-se na Figura 6 que, nos anos 2011 a 2013, houve uma frequência de palavras-chaves como Markowitz e modelo média-variância. Já no período de 2019 a 2021, observa-se o aumento da ocorrência de palavras-chaves como "machine learning", "redes neurais", "lógica Fuzzy" e "K-means", "*multi agent systems selection basead*". Todas essas palavras estão ligadas a métodos que utilizam como base a inteligência artificial. Com o intuito de analisar os estudos mais relevantes nos últimos anos que utilizaram diferentes métodos, foi elaborada a Tabela 5, que apresenta os 5 estudos mais relevantes nos últimos 10 anos.

Tabela 5- Top 5 estudos que possuem mais citações nos últimos 10 anos

| Título | Autor | Citações |
|---|---|---|
| A hybrid stock selection model using genetic algorithms and support vector regression | Huang (2012) | 157 |
| Project selection in project portfolio management: An artificial neural network model based on critical success factors | Constantino, Gravio e Nonino (2015) | 105 |
| Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization | Deng, Li e Lo (2012) | 92 |
| Better than dynamic mean-variance: time inconsistency and free cash flow stream | Li, Wang e Zhou (2012) | 92 |
| Constrained Portfolio Selection using Particle Swarm Optimization | Golmakani e Fazel (2011) | 88 |

Fonte: Elaborado pelos autores com os dados da pesquisa.

O primeiro estudo, com 157 citações, é de Huang (2012) e utiliza um método híbrido que combina algoritmos genéticos e regressão vetor suporte para estimar os retornos das ações, ou seja, faz uso de métodos de *machine learning*, o que vai de encontro com a Figura 6. O segundo estudo da lista, Costantino; Di Gravio e Nonino (2015), faz uso de redes neurais e outros modelos de inteligência artificial para otimizar e selecionar um portfólio. O terceiro artigo, de Deng; Lin e Lo, (2012), também utilizou técnicas de *machine learning* combinadas com o modelo média-variância e métodos heurísticos de otimização de exame de partículas. O quarto artigo, que tem como um dos autores Li X, citado na Tabela 3, utiliza o modelo média-variância combinado com programação dinâmica (CUI et al., 2012). O quinto estudo, de Golmakani e Fazel, (2011), inclui quatro restrições

ao modelo média –variância com o intuito de melhorar o desempenho do portfólio. Após realizar a análise, gera-se um segundo portfólio pela otimização de particular e algoritmo genético.

## 4 Considerações Finais

O objetivo deste capítulo foi realizar uma análise bibliométrica na área de otimização de portfólio de investimentos, com a finalidade de identificar os principais estudos sobre o tema bem como as lacunas e tendências de pesquisas. Para isso, foram coletados dados da base *Scopus*, contemplando o período de 1970 a 2022, sendo os dados tratados pelo pacote Bibliometrix da linguagem de programação R.

Inicialmente, foi realizado o levantamento do número de publicações ao longo dos anos, o que indica um crescimento considerado expressivo da área. Também foi feito o levantamento dos principais autores e suas localidades, tendo sido identificada uma predominância da China e Estados Unidos. Foram listados os periódicos mais relevantes na área, considerando os índices H, G e M, além do número de citações e publicações. Por fim, foi feita uma análise de palavras-chaves e tendências da área, a partir da qual, especificamente, notou-se uma crescente aplicação de Métodos utilizando Lógica Fuzzy e Modelos de *machine learning* e alguns métodos de *deep learning*. Podemos observar a vasta possibilidade de se utilizar diferentes métodos para otimizar portfólios de investimentos, principalmente devido à aplicação de modelos advindos da inteligência artificial. Estes podem tanto ser utilizados individualmente quanto combinados com o tradicional modelo média-variância proposto por Markowitz (1952).

A revisão bibliométrica realizada neste estudo traz as seguintes conclusões i) o campo de estudo está em pleno crescimento nos últimos anos, conforme visto na Figura 2 ; ii) o número de pesquisas utilizando métodos advindos da inteligência artificial como *machine learning* e *deep learning* tem aumentado consideravelmente, conforme visto na Figura 6; iii ) O modelo média – variância de Markowitz (1952) ainda é muito utilizado, porém, há um crescente número de pesquisas que combina o modelo tradicional com outros métodos de otimização (Ballings et al., 2015; Chen, 2021; Ma et al., 2021; Matías and Reboredo, 2012).

Salienta-se que a pesquisa atingiu o objetivo proposto de realizar uma análise bibliométrica acerca de otimização de portfólio de investimentos, porém, ressalta-se como limitações da pesquisa a possibilidade de ter considerado uma segunda base de dados como a *Web of Science*.

## REFERÊNCIAS

ARIA, M.; CUCCURULLO, C. bibliometrix : An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959–975, nov. 2017.

AVRAMOV, D. Stock return predictability and model uncertainty. **Journal of Financial Economics**, v. 64, n. 3, p. 423–458, jun. 2002.

BALLINGS, M. et al. Evaluating multiple classifiers for stock price direction prediction. **Expert Systems with Applications**, v. 42, n. 20, p. 7046–7056, 2015.

CHEN, Y. BP Neural Network Based on Simulated Annealing Algorithm Optimization for Financial Crisis Dynamic Early Warning Model. **Computational Intelligence and Neuroscience**, v. 2021, 2021.

CORBERÁN-VALLET, A. et al. A new approach to portfolio selection based on forecasting. **Expert Systems with Applications**, v. 215, 2023.

COSTA, D. F. et al. Bibliometric analysis on the association between behavioral finance and decision making with cognitive biases such as overconfidence, anchoring effect and confirmation bias. **Scientometrics**, v. 111, n. 3, p. 1775–1799, 2017.

COSTANTINO, F.; DI GRAVIO, G.; NONINO, F. Project selection in project portfolio management: An artificial neural network model based on critical success factors. **International Journal of Project Management**, v. 33, n. 8, p. 1744–1754, nov. 2015.

CUI, X. et al. Better than dynamic mean-variance: Time inconsistency and free cash flow stream. **Mathematical Finance**, v. 22, n. 2, p. 346–378, 2012a.

CUI, X. et al. BETTER THAN DYNAMIC MEAN-VARIANCE: TIME INCONSISTENCY AND FREE CASH FLOW STREAM. **Mathematical Finance**, v. 22, n. 2, p. 346–378, abr. 2012b.

DEMIGUEL, V. et al. A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. **Management Science**, v. 55, n. 5, p. 798–812, maio 2009.

DENG, G.-F.; LIN, W.-T.; LO, C.-C. Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization. **Expert Systems with Applications**, v. 39, n. 4, p. 4558–4566, mar. 2012.

DREZE, J. H. Market allocation under uncertainty. **European Economic Review**, v. 2, n. 2, p. 133–165, dez. 1970.

FAN, J.; FAN, Y.; LV, J. High dimensional covariance matrix estimation using a factor model. **Journal of Econometrics**, v. 147, n. 1, p. 186–197, nov. 2008.

GE, H. et al. Equilibrium strategy for a multi-period weighted mean-variance portfolio selection in a Markov regime-switching market with uncertain time-horizon and a stochastic cash flow. **Communications in Statistics - Theory and Methods**, p. 1–36, 25 ago. 2021.

GHAOUI, L. EL; OKS, M.; OUSTRY, F. Worst-Case Value-At-Risk and Robust Portfolio Optimization: A Conic Programming Approach. **Operations Research**, v. 51, n. 4, p. 543–556, ago. 2003.

GOLMAKANI, H. R.; FAZEL, M. Constrained portfolio selection using particle swarm optimization. **Expert Systems with Applications**, v. 38, n. 7, p. 8327–8335, 2011a.

GOLMAKANI, H. R.; FAZEL, M. Constrained Portfolio Selection using Particle Swarm Optimization. **Expert Systems with Applications**, v. 38, n. 7, p. 8327–8335, jul. 2011b.

HUANG, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. **Applied Soft Computing**, v. 12, n. 2, p. 807–818, fev. 2012.

KALAYCI, C. B.; ERTENLICE, O.; AKBAY, M. A. A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. **Expert Systems with Applications**, v. 125, p. 345–368, 2019.

KOLM, P. N.; TÜTÜNCÜ, R.; FABOZZI, F. J. 60 Years of portfolio optimization: Practical challenges and current trends. **European Journal of Operational Research**, v. 234, n. 2, p. 356–371, 2014.

LEVY, H.; SARNAT, M. The Portfolio Analysis of Multiperiod Capital Investment Under Conditions of Risk. **The Engineering Economist**, v. 16, n. 1, p. 1–20, 24 jan. 1970.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, 2021.

MARKOWITZ, H. Portfolio selection. **The Journal of Finance**, v. 7, n. 1, p. 77–91, 1952.

MATÍAS, J. M.; REBOREDO, J. C. Forecasting performance of nonlinear models for intraday stock returns. **Journal of Forecasting**, v. 31, n. 2, p. 172–188, 2012.

MEHLAWAT, M. K. Credibilistic mean-entropy models for multi-period portfolio selection with multi-choice aspiration levels. **Information Sciences**, v. 345, p. 9–26, 2016.

PALTRINIERI, A. et al. A bibliometric review of sukuk literature. **International Review of Economics & Finance**, abr. 2019.

PATTNAIK, D. et al. Trade credit research before and after the global financial crisis of 2008 – A bibliometric overview. **Research in International Business and Finance**, v. 54, p. 101287, dez. 2020.

ROCKAFELLAR, R. T.; URYASEV, S. Conditional value-at-risk for general loss distributions. **Journal of Banking & Finance**, v. 26, n. 7, p. 1443–1471, jul. 2002.

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

UTZ, S. et al. Tri-criterion inverse portfolio optimization with application to socially responsible mutual funds. **European Journal of Operational Research**, v. 234, n. 2, p. 491–498, 2014.

ZHOU, Z. et al. Big data and portfolio optimization: A novel approach integrating DEA with multiple data sources. **Omega**, v. 104, p. 102479, out. 2021.

## PRODUTO 2 (Bibliográfico): SELEÇÃO DE ATIVOS E OTIMIZAÇÃO DE PORTFÓLIOS DE INVESTIMENTOS COM MÉTODOS DE INTELIGÊNCIA ARTIFICIAL:UMA REVISÃO SISTEMÁTICA E BIBLIOMÉTRICA DA LITERATURA

**Resumo:** O objetivo desta pesquisa é realizar a análise bibliométrica relacionada aos métodos de inteligência artificial aplicados na seleção de ativos e na otimização de portfólio de investimentos. A pesquisa foi realizada no período de 1989 a 2022 e a busca principal retornou um total de 630 artigos na Web of Science e 847 artigos na base Scopus, tendo contemplado todo o período de busca disponível em ambas as bases. As duas bases de dados tiveram os dados consolidados em um único arquivo, sendo removidos os artigos duplicados, restando assim um total de 1.158 artigos. Os dados foram tratados pelo pacote Bibliometrix da linguagem R. No primeiro momento, realizou-se uma análise do número de publicações ao longo dos anos, em que se pode observar um crescimento de 22 % da média de publicações dos últimos 5 anos em relação à média geral. Também foi realizada uma análise dos estudos mais citados, que mostram uma unanimidade na aplicação de métodos advindos da inteligência artificial, tanto para seleção quanto para otimização de portfolio. Por fim, foi feita uma análise de tendências da área, tendo se destacado estudos que utilizaram métodos de *machine learning* ou *deep learning* para realizar a pré-seleção de ações com base em indicadores financeiros para formação de portfólio antes da otimização.

**Palavras-chave**: *deep learning*, *machine learning*, Bibliometrix, pré-seleção.

## Title: ASSET SELECTION AND OPTIMIZATION OF INVESTMENT PORTFOLIO WITH ARTIFICIAL INTELLIGENCE METHODS: A SYSTEMATIC AND BIBLIOMETRIC REVIEW OF THE LITERATURE

Abstract : The aim of this research is to carry out a bibliometric analysis related to artificial intelligence methods applied in the selection and optimization of investment portfolios. The research was carried out from 1989 to 2022 and the main search returned a total of 630 articles in the Web of Science and 847 articles in the Scopus database, covering the entire search period available in both databases. The two databases had their data consolidated into a single file, removing duplicate articles, thus leaving a total of 1.158 articles. The data were treated by the Bibliometrix package of the R language. At first, an analysis of the number of publications over the years was carried out, where it can be observed that there was a growth of 22% in the average of publications in the last 5 years in relation to the average general. Analyzes of the most cited studies were also carried out, which show unanimity in the application of methods derived from artificial intelligence, for selection and for portfolio optimization. Finally, a trend analysis of the area was carried out, highlighting studies that used machine learning or deep learning methods to perform the pre-selection of stocks based on financial indicators, for portfolio formation before optimization.

**Keywords:** *deep learning,* machine learning, Bibliometrix.

## 1 INTRODUÇÃO

Na década de 50, Markowitz (1952) propôs a teoria moderna de portfólio que, por meio da diversificação de ativos e da análise de risco, retorno e correlação, buscou minimizar o risco do portfólio a partir de um determinado retorno desejado. O método proposto por Markowitz (1952) ficou conhecido como modelo de média-variância. Desde então, diversos estudos têm aplicado o modelo de média-variância na composição de carteiras de investimentos (Kalayci et al., 2019). No entanto, estudos (H.R. Golmakani and Fazel, 2011; Mehlawat, 2016) têm utilizado novos métodos na busca de otimizar o portfólio, sendo que alguns (Utz et al., 2014) combinam o modelo de média-variância (Markowitz, 1952) com outros métodos quantitativos, oriundos da estatística, pesquisa operacional, séries temporais e computação científica, cujo intuito é o de obter desempenhos considerados melhores na relação risco e retorno.

Outros métodos que ganharam espaço ao longo dos últimos anos são os modelos provenientes da inteligência artificial: *Machine Learning* (aprendizado de máquina) *e Deep Learning* (aprendizado profundo), que podem ser combinados com o modelo média –variância para melhorar o desempenho do portfólio (Chen, 2021; Ma et al., 2021). Dentre os estudos que fazem uso desses métodos advindos da inteligência artificial, pode-se citar o algoritmo *random forest* (floresta aleatória), que foi utilizado na otimização de portfólio e pré-seleções de ações (Ballings et al., 2015), bem como a regressão de vetor suporte, que foi aplicada ao mercado de ações (Matías and Reboredo, 2012). Os algoritmos de aprendizado de conjunto incluem principalmente *Adaptive Boosting* (AdaBoost) (Zhang et al., 2016), *Gradient Boosted Decision Tree* (GBDT) (Zhou et al., 2019) e *eXtreme Gradient Boosting* (XGBoost) (Nobre and Neves, 2019).

Diante desse contexto, este trabalho visa responder a seguinte pergunta de pesquisa: Quais são os principais métodos de inteligência artificial aplicados na área de seleção e otimização de portfólio de investimentos? Assim, o objetivo do presente capítulo é realizar uma análise bibliométrica na área de seleção e otimização de portfólio de investimentos utilizando métodos de inteligência artificial. Também, espera-se identificar os principais estudos sobre o tema bem como as lacunas e tendências de pesquisas na literatura relacionada

### 2.2.1 Metodologia

A pesquisa caracteriza-se como uma análise bibliométrica, que foi realizada utilizando os artigos extraídos da base *Scopus* e *Web of Science*, sendo também classificada como descritiva e dotada de abordagem quantitativa (Costa et al., 2017). Para realizar a análise bibliométrica, foram seguidas as etapas listadas na Figura 1.

Figura 1- Etapas da montagem da pesquisa e análise bibliométrica.


Fonte: Elaborado pelos autores - adaptado a partir de (Costa et al., 2017).

Observam-se na Figura 1 as etapas que foram seguidas no decorrer da pesquisa. A primeira etapa de delimitar o objetivo e o campo da pesquisa ocorreu na introdução do trabalho. As bases de dados que serão utilizadas para realizar o estudo serão a base *Scopus* e *Web of Science*. A metodologia aplicada nas etapas subsequentes será descrita nos próximos tópicos.

### 2.2.1.1 Procedimentos Metodológicos

Para a definição e identificação dos termos de busca e palavras-chaves, considerando a temática das técnicas de inteligência artificial ligadas à otimização de portfólio de investimentos, foram utilizadas expressões booleanas com o intuito de trazer à tona o maior número de trabalhos que contemplam a temática. A chave de pesquisa utilizada foi:

(PORTFOLIO AND (OPTIMIZATION OR SELECTION ) AND ( "ARTIFICIAL INTELLIGENCE " OR " MACHINE LEARNING" OR "DEEP LEARNING" ) AND (INVESTMENT OR FINANCE)

### 2.2.1.2 Coleta e estruturação de dados

Na terceira etapa da Figura 1, coleta e estruturação de dados, para análise bibliométrica foi utilizado o pacote *Bibliometrix* do *software R*, programa utilizado na literatura conforme estudos (Khan, 2022; Paltrinieri et al., 2019; Pattnaik et al., 2020) pela praticidade de quantificar e processar os dados. Para realizar as análises utilizando o *Bibliometix*, seguiu-se o estudo de (Aria and Cuccurullo, 2017), que apresenta uma série de possibilidades de análises bibliométricas com o pacote e suas respectivas funções.

#### 2.2.1.3 Análise contextual da produção científica relacionada a amostra

A quarta etapa da Figura 1 consiste em realizar uma análise contextual da produção científica selecionada pela amostra cujo objetivo é conhecer as publicações selecionadas pela busca. Para tanto, primeiramente foi realizada uma análise do número de publicações ao longo dos anos. Além disso, também foram analisadas as citações dos artigos selecionados, a fim de identificar as publicações mais relevantes da amostra. Também foi realizada uma análise sobre os periódicos mais relevantes do tema e os países que têm produzido as pesquisas mais impactantes ao longo do tema.

#### 2.2.1.4 Análise das redes das citações realizadas pela amostra

A quinta etapa da análise bibliométrica, na Figura 1, consiste em analisar e discutir as redes de citações realizadas pela amostra. Essa etapa tem por objetivo identificar os *clusters* de pesquisas e as relações de citações.

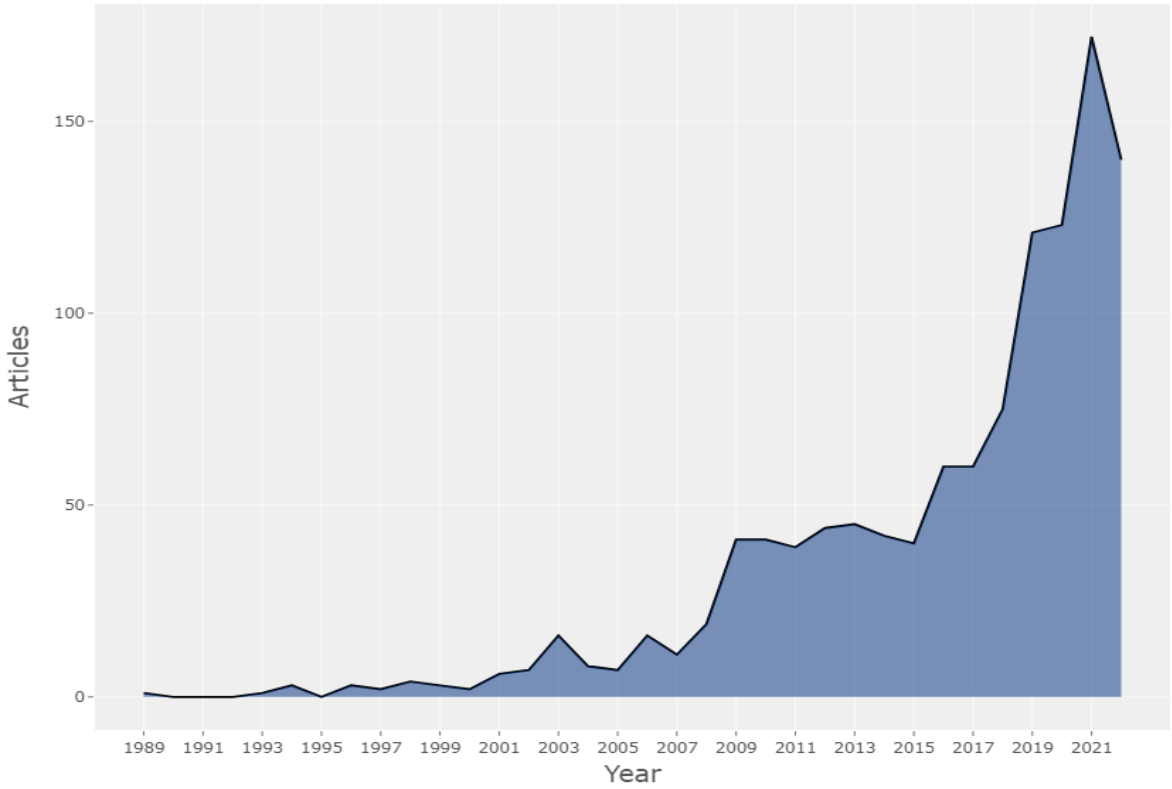
## 3Resultados e discussões

### 3.1 Análise contextual da amostra e da produção científica

A pesquisa foi realizada em outubro de 2022 e a busca principal retornou um total de 630 artigos na *Web of Science* e 847 artigos na base *Scopus*, que foram publicados no período de 1989 a 2022. A pesquisa contemplou todo o período de busca disponível em ambas as bases. As duas bases de dados tiveram os dados consolidadas em um único arquivo, sendo removidos os artigos duplicados, restando assim um total de 1158 artigos.

Com o objetivo de conhecer o recorte temporal das publicações recolhidas, elaborou-se um gráfico a partir de um critério de busca previamente estabelecido, que pode ser observado na Figura 2, mostrando o número de artigos publicados ao longo do tempo.

Figura 2: Análise das publicações ordenadas por quantidade anual de publicações




Fonte: Elaborada pelos autores com os dados da pesquisa.

Observa-se na Figura 2 um crescimento no número de publicações principalmente a partir de 2009. O primeiro artigo da amostra foi publicado em 1989, intitulado “Intelligent Stock Portfolio Management System”, publicado na revista “Expert Systems”, tendo 14 citações na base *Scopus*. O estudo de Lee; Kim; Chu, (1989) propõe um modelo de seleção e otimização de carteira utilizando algoritmos inteligentes.

Dentre os artigos selecionados com base nos critérios de busca previamente estabelecidos, alguns se destacaram de acordo com o número de citações. Portanto, a Tabela 1 apresenta os estudos com mais citações conforme a busca realizada.

Tabela 1- Cinco artigos mais citados

| Título | Autor | Citações |
|---|---|---|
| Computational Intelligence and Financial Markets: A Survey and Future Directions | (Cavalcante et al., 2016) | 245 |
| Portfolio selection using neural networks | (Fernández and Gómez, 2007) | 222 |
| A first application of independent component analysis to extracting structure from stock returns | (Back and Weigend, 1997) | 199 |

| | | |
|---|---|---|
| Particle Swarm Optimization (PSO) for the constrained portfolio optimization problem | (Zhu et al., 2011) | 181 |
| A hybrid stock selection model using genetic algorithms and support vector regression | (Huang, 2012a) | 161 |

Fonte: Elaborada pelos autores com os dados da pesquisa.

Com 245 citações, Cavalcante et al., (2016) apresentam uma revisão sistemática dos métodos de inteligência computacional aplicados à área de finanças, tanto na previsão de retornos de preços de ações como na seleção e otimização de portfólios de investimentos. O segundo artigo apresentado na Tabela 1, com 222 citações, desenvolveu um modelo heurístico que faz uso de redes neurais artificiais para selecionar e otimizar portfólios de investimentos (Fernández and Gómez, 2007). Com 199 citações, o estudo de Back; Weigend (1997) é o terceiro estudo com maior número de citações, sendo que a pesquisa dos autores utiliza uma técnica de processamento de linguagem de sinais conhecida como análise de componentes independentes, sendo aplicada em séries financeiras multivariadas como a carteira de investimento. O artigo de Zhu et al. (2011), que tem 181 citações, apresenta a aplicação do modelo de otimização de enxame por partículas para otimizar um portfólio de investimentos e compara os resultados com outros métodos. O estudo de Huang (2012), com 161 citações, utilizou um modelo híbrido que combina dois métodos de inteligência artificial para seleção de ativos: regressão vetor suporte e algoritmo genético.

A pesquisa também analisou os periódicos com maior relevância no tema, sendo considerados os índices H, G e M, TC (total de citações e NP (número de publicações. A Tabela 2 apresenta os periódicos com maior impacto, levando em consideração esses parâmetros.

Tabela 2- Dez periódicos mais ranqueados conforme o número de citações

| Periódicos | h_index | g_index | m_index | TC | NP |
|---|---|---|---|---|---|
| EXPERT SYSTEMS WITH APPLICATIONS | 19 | 40 | 1,12 | 1669 | 42 |
| EUROPEAN JOURNAL OF OPERATIONAL RESEARCH | 7 | 11 | 0,27 | 657 | 11 |
| QUANTITATIVE FINANCE | 6 | 11 | 0,75 | 148 | 11 |
| KNOWLEDGE-BASED SYSTEMS | 6 | 10 | 0,67 | 660 | 10 |

| | | | | | |
|---|---|---|---|---|---|
| ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING | 2 | 5 | 0,25 | 29 | 7 |
| COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE | 1 | 2 | 0,08 | 6 | 7 |
| INFORMATION SCIENCES | 6 | 7 | 0,38 | 254 | 7 |
| JOURNAL OF FINANCIAL DATA SCIENCE | 4 | 5 | 1,00 | 25 | 7 |
| APPLIED SOFT COMPUTING JOURNAL | 5 | 6 | 0,45 | 243 | 6 |
| INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS | 3 | 4 | 0,30 | 18 | 6 |

Fonte: Elaborada pelos autores com os dados da pesquisa.

### 3.2 Análise das palavras-chaves e tendências de pesquisas

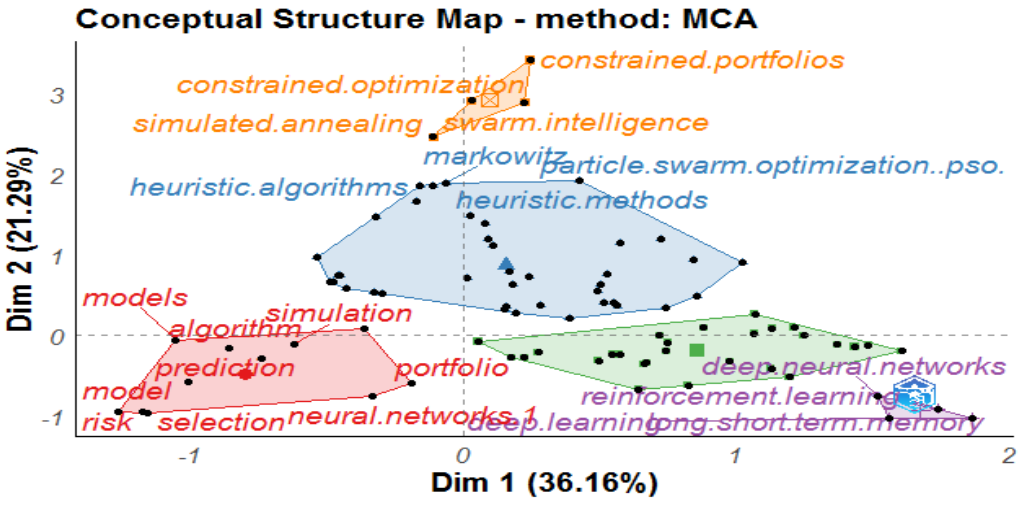
A Figura 3 apresenta neste tópico as *Keywords* (palavras-chaves) que apareceram com mais frequências na amostra do presente estudo, a partir da base *Scopus* e *Web of Science*. Em destaque, podemos notar os termos "investimentos", "mercado de finanças", "*machine learning*" e "neural networks", dentre outras obtidas, conforme apresentado na Figura 9.

Figura 3: Nuvem de palavras-chaves


Fonte: Elaborado pelos autores com os dados da pesquisa.

Na Figura 3, observa-se a ocorrência principal de palavras-chaves como *"machine learning*" e *"investiment*".Um ponto a se destacar é que esses termos estão ligados à chave de busca que foi determinada na seção de procedimentos metodológicos. Outras palavras-chaves que apareceram com frequência e merecem destaque são: *"neurol networks"*ou rede neural, que está ligada a uma subárea da inteligência artificial. Outra palavra que merece destaque é "*Computer simulation*", que diz respeito à simulação computacional na área de otimização.

Figura 4: Nuvem de palavras-chaves



Fonte: Elaborado pelos autores com os dados da pesquisa.

A Figura 4 apresenta um dendrograma que utiliza técnicas de agrupamento para tentar identificar grupos de artigos que expressam objetivos em comum. Nesse caso, pode-se relatar quatro grupos.O grupo em vermelho trata de estudos sobre seleção de ativos por meio de técnicas de previsão de retorno usando *machine learning* e redes neurais, a exemplo dos estudos (Ma et al., 2021; Prasad and Bakhshi, 2022). O grupo em azul ressalta estudos que utilizaram na otimização métodos heurísticos e algoritmos bio-inspirados como exame de partículas e algoritmo genético, a exemplo do estudo de (Hamid Reza Golmakani and Fazel, 2011) e o estudo de (Frausto Solis et al., 2022), que utilizam algoritmo genético tanto para seleção de investimentos como para otimização do modelo. Outra dimensão em laranja trata de diferentes métodos de otimização que buscam encontrar as melhores soluções ótimas, a exemplo dos estudos que utilizam o algoritmo *simulated annealing* (Kirkpatrick et al., 1983; Luo et al., 2014; Wang et al., 2016), e portfólio que fazem uso de restrições de cardinalidades (Deng et al., 2022; Shahid et al., 2022).

Por fim, a última dimensão trata de modelos que utilizam *deep learning* e *deep reforcing learning*, dentre os quais se pode destacar estudos que utilizaram redes neurais para fazer a pré-seleção de ativos para uma posterior otimização (Michańków et al., 2022; Padhi et al., 2022a). Uma outra tendência relatada nos estudos que utilizaram métodos de inteligência artificial, seja de *machine learning* ou *deep learning*, é que quando a pré-seleção de ações é realizada por tais métodos antes da otimização, o portfólio apresenta um desempenho largamente superior se comparado com outros portfólios(Chen et al., 2021; Ma et al., 2021) .

## 4. Considerações Finais

O objetivo deste trabalho foi realizar uma análise bibliométrica na área de otimização de portfólio de investimentos utilizando inteligência artificial, com a finalidade de identificar os principais estudos sobre o tema, bem como as lacunas e tendências de pesquisas. Para isso, foram coletados dados da base *Scopus* e *Web of Science*, contemplando todo o período de dados disponível, sendo os dados tratados pelo pacote Bibliometrix da linguagem de programação R.

Inicialmente, foi realizado o levantamento do número de publicações ao longo dos anos, o que indica um crescimento considerado expressivo da área. Também foi feito o levantamento dos principais autores e suas localidades, tendo sido identificada uma predominância da China e Estados Unidos. Foram listados os periódicos mais relevantes na área, considerando os índices H, G e M, além do número de citações e publicações. Por fim, foi feita uma análise de palavras-chaves e tendências da área, a partir da qual, especificamente, notou-se uma crescente aplicação de métodos utilizando de *deep learning e deep refoircening learning.* Podemos observar a vasta possibilidade

de se utilizar diversos métodos, tanto para realizar a pré-seleção de ativos quanto para otimizar portfólios de investimentos, principalmente devido à aplicação de modelos advindos da inteligência artificial. Estes podem tanto ser utilizados individualmente quanto combinados com o tradicional modelo média-variância proposto por (Markowitz, 1952).

A revisão bibliométrica realizada neste estudo traz as seguintes conclusões: i) o campo de estudo está em pleno crescimento nos últimos anos, conforme visto na Figura 8 ; ii) o número de pesquisas utilizando métodos advindos da inteligência artificial como *machine learning* e *deep learning* tem aumentado consideravelmente, conforme visto na Figura 10; iii ) O modelo média – variância de Markowitz (1952) ainda é muito utilizado, porém, há um grande número de pesquisas que combina o modelo tradicional com outros métodos (Ballings et al., 2015; Chen et al., 2021; Ma et al., 2021; Matías and Reboredo, 2012); iv) alguns estudos constataram que a aplicação de métodos de inteligência artificial para realizar a seleção de ações antes de realizar a otimização do portfólio aumentou significativamente o desempenho do portfólio (Ma et al., 2021; Ta et al., 2020; Zhou et al., 2019).

Para futuras pesquisas, sugere-se: i) realizar um estudo das técnicas de inteligência artificial focadas na pré-seleção de ativos; ii) realizar uma análise dos métodos aplicados à temática e relacionar com os resultados encontrados nas pesquisas; e iii) realizar trabalhos separados pelas áreas da inteligência artificial, por exemplo, "*machine learning* e seleção e otimização de portfolios".

## REFERÊNCIAS


ARIA, M.; CUCCURULLO, C. bibliometrix : An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959–975, nov. 2017.

BACK, A. D.; WEIGEND, A. S. A First Application of Independent Component Analysis to Extracting Structure from Stock Returns. **International Journal of Neural Systems**, v. 08, n. 04, p. 473–484, 21 ago. 1997.

BALLINGS, M. et al. Evaluating multiple classifiers for stock price direction prediction. **Expert Systems with Applications**, v. 42, n. 20, p. 7046–7056, 2015.

CAVALCANTE, R. C. et al. Computational Intelligence and Financial Markets: A Survey and Future Directions. **Expert Systems with Applications**, v. 55, p. 194–211, ago. 2016.

CHEN, W. et al. Mean–variance portfolio optimization using machine learning-based stock price prediction. **Applied Soft Computing**, v. 100, p. 106943, mar. 2021.

CHEN, Y. BP Neural Network Based on Simulated Annealing Algorithm Optimization for Financial Crisis Dynamic Early Warning Model. **Computational Intelligence and Neuroscience**, v. 2021, 2021.

COSTA, D. F. et al. Bibliometric analysis on the association between behavioral finance and decision making with cognitive biases such as overconfidence, anchoring effect and confirmation bias. **Scientometrics**, v. 111, n. 3, p. 1775–1799, 2017.

DENG, X. et al. Non-dominated sorting genetic algorithm-II for possibilistic mean-semiabsolute deviation-Yager entropy portfolio model with complex real-world constraints. **Mathematics and Computers in Simulation**, v. 202, p. 59–78, dez. 2022.

FERNÁNDEZ, A.; GÓMEZ, S. Portfolio selection using neural networks. **Computers & Operations Research**, v. 34, n. 4, p. 1177–1191, abr. 2007.

FRAUSTO SOLIS, J. et al. SAIPO-TAIPO and Genetic Algorithms for Investment Portfolios. **Axioms**, v. 11, n. 2, 2022.

GOLMAKANI, H. R.; FAZEL, M. Constrained portfolio selection using particle swarm optimization. **Expert Systems with Applications**, v. 38, n. 7, p. 8327–8335, 2011a.

GOLMAKANI, H. R.; FAZEL, M. Constrained Portfolio Selection using Particle Swarm Optimization. **Expert Systems with Applications**, v. 38, n. 7, p. 8327–8335, jul. 2011b.

HUANG, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. **Applied Soft Computing**, v. 12, n. 2, p. 807–818, fev. 2012.

KALAYCI, C. B.; ERTENLICE, O.; AKBAY, M. A. A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. **Expert Systems with Applications**, v. 125, p. 345–368, 2019.

KHAN, M. A. ESG disclosure and Firm performance: A bibliometric and meta analysis. **Research in International Business and Finance**, v. 61, p. 101668, out. 2022.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by Simulated Annealing. **Science**, v. 220, n. 4598, p. 671–680, 13 maio 1983.

LEE, J. K.; KIM, H. S.; CHU, S. C. Intelligent Stock Portfolio Management System. **Expert Systems**, v. 6, n. 2, p. 74–87, abr. 1989.

LUO, Y.; ZHU, B.; TANG, Y. Simulated annealing algorithm for optimal capital growth. **Physica A: Statistical Mechanics and its Applications**, v. 408, p. 10–18, 2014.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, 2021.

MARKOWITZ, H. Portfolio selection. **The Journal of Finance**, v. 7, n. 1, p. 77–91, 1952.

MATÍAS, J. M.; REBOREDO, J. C. Forecasting performance of nonlinear models for intraday stock returns. **Journal of Forecasting**, v. 31, n. 2, p. 172–188, 2012.

MEHLAWAT, M. K. Credibilistic mean-entropy models for multi-period portfolio selection with multi-choice aspiration levels. **Information Sciences**, v. 345, p. 9–26, 2016.

MICHAŃKÓW, J.; SAKOWSKI, P.; ŚLEPACZUK, R. LSTM in Algorithmic Investment Strategies on BTC and S&P500 Index. **Sensors**, v. 22, n. 3, 2022.

NOBRE, J.; NEVES, R. F. Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. **Expert Systems with Applications**, v. 125, p. 181–194, jul. 2019.

PADHI, D. K. et al. An Intelligent Fusion Model with Portfolio Selection and Machine Learning for Stock Market Prediction. **Computational Intelligence and Neuroscience**, v. 2022, 2022.

PALTRINIERI, A. et al. A bibliometric review of sukuk literature. **International Review of Economics & Finance**, abr. 2019.

PATTNAIK, D. et al. Trade credit research before and after the global financial crisis of 2008 – A bibliometric overview. **Research in International Business and Finance**, v. 54, p. 101287, dez. 2020.

PRASAD, A.; BAKHSHI, P. Forecasting the Direction of Daily Changes in the India VIX Index Using Machine Learning. **Journal of Risk and Financial Management**, v. 15, n. 12, 2022.

SHAHID, M. et al. Solving constrained portfolio optimization model using stochastic fractal search approach. **International Journal of Intelligent Computing and Cybernetics**, 2022.

TA, V.-D.; LIU, C.-M.; TADESSE, D. A. Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. **Applied Sciences (Switzerland)**, v. 10, n. 2, 2020.

UTZ, S. et al. Tri-criterion inverse portfolio optimization with application to socially responsible mutual funds. **European Journal of Operational Research**, v. 234, n. 2, p. 491–498, 2014.

WANG, X.; HE, L.; JI, H. **Modified generalized simulated annealing algorithm used in data driven portfolio management**. Proceedings of 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2016. **Anais**...2016.

ZHANG, X.; LI, A.; PAN, R. Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine. **Applied Soft Computing**, v. 49, p. 385–398, dez. 2016.

ZHOU, F. et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. **Applied Soft Computing**, v. 84, p. 105747, nov. 2019.

ZHU, H. et al. Particle Swarm Optimization (PSO) for the constrained portfolio optimization problem. **Expert Systems with Applications**, v. 38, n. 8, p. 10161–10169, ago. 2011.

## PRODUTO 3 (Bibliográfico): ANÁLISE DE DESEMPENHO DE PORTFÓLIOS QUE COMBINAM MODELOS DE *MACHINE LEARNING* E OTIMIZAÇÃO MULTIOBJETIVO NO MERCADO DE AÇÕES BRASILEIRO

**Resumo:** Este trabalho objetiva avaliar o desempenho de portfólios compostos por ativos pré-selecionados por meio dos modelos de *machine learning* e posteriormente otimizados pelo modelo multiobjetivo. A metodologia envolve a aplicação dos algoritmos de *machine learning Random Forest* (RF), *eXtreme Gradient Boosting* (XGBoost) e *Multilayer Perceptron* (MLP) para realizar a pré-seleção de ações. Essa seleção é baseada em indicadores técnicos pré-definidos após uma revisão da literatura relacionada e filtrados por meio do método de *Feature Selection Extraclassifica*. O período de análise e coleta de dados foi de janeiro de 2013 a 2023. Os dados dos modelos de *machine learning* foram treinados utilizando validação cruzada e avaliados por meio de métricas como *recall*, precisão, acurácia e F1 *Score*. Após o treinamento dos modelos, cada portfólio passou por uma otimização multiobjetivo que visa maximizar o retorno e minimizar o risco simultaneamente. Além disso, foram consideradas as restrições relacionadas à alocação máxima de ativos de 15% e foi realizada uma análise sem e com a presença dos custos de transações. Os resultados obtidos foram comparados com o *benchmark* do mercado brasileiro, o IBOV, com carteiras ingênuas e com um portfólio que otimizou todos os ativos presentes na amostra inicial, sem a pré-seleção dos ativos. Os portfólios otimizados se destacaram, superando tanto o *benchmark* de mercado como o portfólio formados pela carteira ingênua. Embora todos os modelos apresentem resultados superiores ao *benchmark* de mercado, a combinação entre o modelo *Random Forest* e a otimização multiobjetivo se destacou como a abordagem considerada mais eficaz ao longo do período analisado, visto que obteve desempenhos superiores com um número reduzido de ativos, o que mostra mais eficiência na composição de portfólios no mercado brasileiro.

**Palavras-chave**: *machine learning*; *portfólio*; *random forest* (RF); otimização multiobjetivo; Brasil.

## TITLE: PERFORMANCE ANALYSIS OF PORTFOLIOS THAT COMBINE MACHINE LEARNING MODELS AND MULTI-OBJECTIVE OPTIMIZATION IN THE BRAZILIAN STOCK MARKET

**Abstract:** This work aims to evaluate the performance of portfolios composed of assets pre-selected using machine learning models and subsequently optimized using the multi-objective model. The methodology involves the application of machine learning algorithmsRandom Forest (RF), eXtreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP) to pre-select actions. This selection is based on pre-defined technical indicators after a review of related literature and filtered through the Extraclassifica Feature Selection method. The data analysis and collection period was from January 2013 to 2023. The data for training and for testing using cross-validation and evaluated using metrics, such as recall, precision, accuracy and F1 Score. After model training, each portfolio underwent multi-objective optimization that aims to maximize return and minimize risk simultaneously. Furthermore, restrictions related to the maximum asset allocation of 15% were considered and an analysis was carried out without and with the presence of transaction costs. The results obtained were compared with the Brazilian market benchmark, IBOV, with naive portfolios and with a portfolio that optimized all assets present in the initial sample, without pre-selection of assets. The optimized portfolios stood out, surpassing both the market benchmark and the portfolio formed by the naive portfolio. Although all models present results superior to the market benchmark, the combination of the Random Forest model and multi-objective optimization stood out as the

approach considered most effective throughout the analyzed period, as it obtained superior performance with a reduced number of assets, which shows more efficiency in the composition of portfolios in the Brazilian market.

**Keywords:** *machine learning*; *portfólio*; *random forest* (RF); multi-objective optimization; Brasil.

## 1 INTRODUÇÃO

A seleção de ações é considerada uma tarefa complexa e desafiadora tanto para investidores quanto para instituições, pois é preciso lidar com grandes quantidades de dados e informações (BODNAR; MAZUR; OKHRIN, 2017). Nesse contexto, a aplicação de modelos de *machine learning* e *deep learning* visa auxilar na tomada de decisão e na otimização de resultados (MA; HAN; WANG, 2021).

Recentemente, estudos têm aplicado modelos de *machine learning* e *deep learning* para realizar a pré-seleção de ações e posteriormente otimizar os portfólios pelo modelo média-variância. Como exemplo, o algoritmo extreme *gradient boosting* (Xgboost) foi implementado com base em variação de retorno, tomando como base ativos pertencentes ao índice de Shangai da China (CHEN et al., 2021). O modelo SVM (*support vector machine*) foi utilizado no Brasil, tendo como valor alvo os ativos que atingissem retorno diário acima de 2%, (PAIVA et al., 2019).

Os modelos SVM, *Random forest* (RF), LSMT, CNN, Arima e *deep learning multilayer perceptron* (DMLP) também foram testados para realizar a pré-seleção de ações na China. Após cada modelo selecionar um conjunto de ativos, a otimização de cada carteira foi realizada pelo modelo bi-objetivo, que busca minimizar o risco e maximizar o retorno. Os modelos tiveram seus resuldados comparados, tendo o modelo RF apresentado melhor resultado (MA; HAN; WANG, 2021). Os modelos *machine learning* RF, AdaBoost, XGBoost, SVR, KNN e ANN foram aplicados para realizar a pré-seleção de ativos, considerando os ativos pertencentes ao índice BSE da Índia, sendo que o modelo Adaboost obteve melhores resultados em comparação aos outros modelos (BEHERA et al., 2023). Após realizar a pré-seleção, os modelos foram otimizados pelo modelo média-variância (BEHERA et al., 2023).

O objetivo do presente estudo é avaliar o desempenho de portfólios compostos por ativos pré-selecionados por meio dos modelos de *machine learning* e posteriormente otimizados pelo modelo multiobjetivo. Especificamente, o trabalho testaos modelos de *machine learning Random Forest* (RF), *extreme gradient boosting* (Xgboost) e *multilayer perceptron* (MLP), para realizar a pré-seleção de ações com base em indicadores técnicos. Após a implementação dos modelos, os ativos selecionados tiveram a definição das alocações por meio de um modelo de otimização multiobjetivo e seus resultados foram validados por indicadores de desempenho de portfólios.

.

## 1.1 Contribuições da Pesquisa

Este estudo busca contribuir para a literatura de maneira empírica, demonstrando que portfólios otimizados pelo método multiobjetivo que utilizam modelos de *machine learning* para a pré-seleção de ativos com base em indicadores técnicos alcançam desempenhos considerados mais eficientes quando comparados com desempenho de portfólios que não utilizam essas técnicas de pré-seleção e otimização verificadas neste estudo. Os portfólios são avaliados por uma série de indicadores de desempenho, como índice de Sharpe, índice de Treynor, coeficiente beta, Alfa de Jensen, prêmio pelo risco, e VAR. Também os comparamos com um *benchmark* de mercado e com portfólios formados por uma carteira considerada ingênua (igualmente ponderada). Essa abordagem enriquece a metodologia adotada neste estudo ao permitir uma avaliação considerada mais completa e que permitiu a comparação do desempenho dos diferentes portfólios. Dessa forma, contribuímos para a literatura relacionada ao fornecer uma análise do desempenho dos portfólios, demonstrando empiricamente como a combinação de pré-seleção de ativos por meio de modelos de *machine learning* e otimização multiobjetivo resulta em portfólios que superam um *benchmark* de mercado, o que é valioso para investidores, visto que a metodologia pode ser implementada em outros mercados.

Sinteticamente, destaca-se as seguintes contribuições do estudo i) Desenvolvimento de uma metodologia que utiliza modelos de *machine learning* para a seleção de investimentos, com o objetivo de superar o *benchmark* de mercado. Os modelos são alimentados com indicadores financeiros selecionados através de um processo de *feature selection* e treinados usando métodos de validação cruzada; ii) Integração desse modelo de seleção de investimentos em um modelo de otimização multiobjetivo; iii) Realização de uma análise que avalia o desempenho dos investimentos com base em diversos indicadores financeiros, incluindo índice de Sharpe, índice de Treynor, coeficiente beta, Alfa de Jensen, prêmio pelo risco e VAR. Além disso, a análise contempla a evolução do retorno ao longo do tempo e a acumulação de riqueza no portfólio; iv) Consideração de diferentes cenários de custos de transação para avaliar a eficácia dos modelos.

Os experimentos demostram que os portfólios que ultilizam a metodologia do estudo superam consideravelmente o *benchmark* de mercado e que os portfólios que ultilizam a pré-seleção de ações têm uma vantagem em relação aos outros portfólios de ter um número reduzido de ativos, o que facilitaá a gestão do portfólio pelo investidor.

## 2. Referencial teórico

### 2.1 Literatura relacionada à aplicação de métodos de *machine learning* em carteiras de investimentos

Os modelos de *machine learning* e *deep learning* têm sido aplicados para realizar a pré-seleção de ações, a exemplo do estudo de (HUANG, 2012), que aplicou a técnica *Support vector regression* baseada em indicadores fundamentalistas para selecionar ativos da Bolsa de Valores de Taiwan. O modelo XGBoost foi aplicado por Chen (2020) juntamente com o modelo média-variância, no qual foram incluídas restrições de cardinalidade e custos de transação. Outro estudo que aplicou o modelo XGBoost para realizar pré-seleção foi realizado por Nobres e Neves (2019), porém em conjunto com um modelo multi-objetivo. Foram utilizados indicadores técnicos como dados de entrada no modelo, os quais foram redimensionados a partir de uma análise de componentes principais.

O estudo de Ma; Han e Wang, (2021) aplicou os modelos de *Deep multilayer perceptron, Long short term memory, Convolutional neural network, Support vector regression, Random forest* para realizar a pré-seleção de ações na bolsa da China e, posteriormente, foram combinados com modelo média-variância, buscando qual modelo traria melhor resultado. Em sua pesquisa, Ma; Han e Wang, (2021) tiveram o portfólio otimizado pelo modelo média-variância. Os ativos selecionados pelo modelo *Random forest* apresentaram melhores resultados em comparação com outros modelos ao longo do período testado.

No Brasil, há dois estudos que aplicaram modelos de machine learning para pré-seleção de ações foram realizados por Paiva et al. (2019) e Silva et al. (2024). Ambos utilizaram o modelo SVM. No estudo de Paiva et al. (2019), foram utilizadas variações de retornos como dados de entrada e o modelo foi combinado com média-variação. Já Silva et al. (2024) utilizaram indicadores fundamentalistas trimestrais como dados de entrada e combinaram com um modelo de otimização que visava maximizar o índice de Sharpe.

A Tabela 1 apresenta um conjunto de estudos mais recentes que utilizaram métodos de *machine learning* e *deep learning* para realizar a pré-seleção de ativos antes da otimização, juntamente com seus respectivos modelos, dados de entrada e o país onde o estudo foi conduzido. Também é indicado se os dados de entrada foram submetidos a um processo de seleção de características ou redução de dimensionalidade, se foram treinados por validação cruzada, qual método de otimização foi aplicado após a seleção, se houve restrições e se o estudo incluiu custos de transação. Além disso, a tabela apresenta uma coluna com a abordagem proposta neste estudo.

Tabela 1- Estudos recentes que utilizaram métodos para realizar pré-seleção de ativos

| Autor | (PAIVA et al., 2019) | (HUANG, 2012) | (CHEN et al., 2021) | Silva et al. (2024) | (MA; HAN; WANG, 2021) | Neves, Nobre (2019) | (WANG et al., 2020) | (CHAWEEWANCHON; CHAYSIRI, 2022) | (MAZRAEH et al., 2022) | (Ashrafzadeh et al. 2023) | (BEHERA et al., 2023) | Abordagen proposta |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo | SVM | SVM | Xgboost | SVM | *DMP, LSM, CNN, SVR, Random forest* | *Xgboost* | LSTM, SVM, Random forest | CNN e BiLSTM | SVM e Random forest | CNN | *Random Forest, XGBoost, AdaBoost, SVR,KNN, ANN* | Randon Forest, Xgboost, MLP |
| Dados de entrada | Variação de retornos e indicadores técnicos | Indicadores funtamentalistas | Variação de retornos | Indicadores fundamentalistas | Variação de retornos | Indicadores técnicos | Indicadores técnicos e variação de retornos | Variação de retornos | Indicadores técnicos | Indicadores técnicos | Variação de retornos | Indicadores técnicos |
| País | Brasil | Tawain | China | Brasil | China | Estados Unidos | Inglaterra | Tailândia | Irã | Estados Unidos | Índia ,Japão e China | Brasil |
| Seleção de caracteristica ou redução de dimensionalidad | - | | - | - | - | PCA | - | - | - | Análise de Clusters | - | Extratreeclassifier |
| Cross Validation | - | | - | - | - | | - | - | - | - | - | x |
| Modelo de Otimização | Cardinalidade | | Média - variancia | Max. Sharpe | Média - variancia | Mul-objetivo | Média - variancia | Média - variancia | Mult-objetivo | Média - variancia | Mult-objetivo | Mult-objetivo |
| Restrições do modelo | x | | Cardinalidade | - | - | - | Cardinalidade | - | - | - | - | Alocação Máxima |
| Custos de transação | | | x | - | x | - | x | - | - | x | x | x |

## 2.2. Modelos de *machine learning*

### 2.2.1 Random Forest (RF)

O algoritmo *Random forest* (RF) é um modelo não paramétrico e não linear que foi proposto por (TIN KAM HO, 1995). Esse algoritmo consiste em criar várias árvores de decisão a partir de uma amostra aleatória dos dados de treino e com variáveis selecionadas de forma aleatória, permitindo que o modelo seja mais robusto e menos sensível a dados ruidosos, evitando assim o problema de *overfitting*, que se caracteriza quando os dados fornecem previsões precisas para dados de treino e não para novos dados (BREIMAN, 2001).

Devido à sua versatilidade, esse modelo tem sido aplicado em problemas de seleção de investimentos, como no estudo para realizar pré-seleção de ações da bolsa da China, em que se mostrou superior aos métodos LSTM e CNN (MA; HAN; WANG, 2021). O algoritmo também foi aplicado para detectar se os preços de ações americanas iriam subir ou diminuir, sinalizando assim os melhores investimentos de compra de ações (BASAK et al., 2019).

### 2.2.2 eXtreme Gradient Boosting (Xgboost)

O algotirmo *eXtreme Gradient Boosting* (Xgboost), proposto por (CHEN; GUESTRIN, 2016), assim como outros algoritmos de *boosting*, faz uso de árvores de decisão para seu modelo de conjunto, sendo cada árvore um aprendiz fraco. O algoritmo segue construindo sequencialmente mais árvores de decisão, cada uma corrigindo o erro da árvore anterior até que uma condição de parada seja alcançada. Diferentes penalidades de regularização são aplicadas para evitar *overfitting*, produzindo um treinamento bem-sucedido para que o modelo possa generalizar adequadamente.

Na área de seleção de investimentos, o Xgboost foi aplicado para selecionar ações de um índice de Xangai, sendo posteriormente combinado com um modelo de otimização de portfólio (CHEN et al., 2021) . Outro estudo aplicou o algoritmo para prever tendência de preços dos ativos pertencentes ao índice S&P 500. A partir das previsões, os autores conseguiram indicar em quais ativos era mais propenso investir (ZHOU et al., 2019).

As características desse método são: baixa complexidade computacional, rápido em velocidade de execução e alta precisão (CHEN et al., 2021). A função objetivo do Xgboost é combinar o termo de penalidade padrão com o termo de função de perda para obter a solução ótima, na qual o termo de penalidade regular reduz a variância do modelo, evitando assim o ajuste excessivo (CHEN; GUESTRIN, 2016).O modelo aditivo do modelo de conjunto de árvores é descrito por:

$$\hat{y}_i = \sum_{k=1}^{k} f_k\,(x_i), f_k \in F \qquad (1)$$

Onde K é o número de árvores, $f_k$ é uma função no espaço funcional F e F é o conjunto de todas as combinações possíveis. A função objetivo a ser otimizada é dada pela equação 2.

$$L(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{K=1}^{K} \omega(f_k) \quad (2)$$

Onde $l(y_i, \hat{y}_i)$ é a função de perda e $\omega(f_k)$ , a chamada função de punição regular, que é calculada com forme equações abaixo:

$$\omega(f_k) = yT + \frac{1}{2}\Lambda\|\omega\|^2 \quad (3)$$

De acordo com de Chen e Guestrin (2016) , a árvore de decisão possui diversos parâmetros. No algoritimo Xgboost, utiliza-se a estratégia aditiva para corrigir o que foi aprendido nas árvores anteriores e é adicionada uma nova árvore de cada vez, conforme as equações abaixo:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0, \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(0)} + f_2(x_i) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_1(x_i) \end{aligned} \quad (4)$$

Após a aplicação das equações acima, a função objetivo do modelo passa a ser:

$$\hat{L}^t = \sum_{i}^{T} l((y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \omega(f_k) \quad (5)$$

A equação passa pela expansão de Taylor e por duas derivadas, chegando à equação 6:

:

$$\hat{L}^t \approx \sum_{i}^{T} [l((y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i {f_t}^2(x_i)] + \omega(f_t) \quad (6)$$

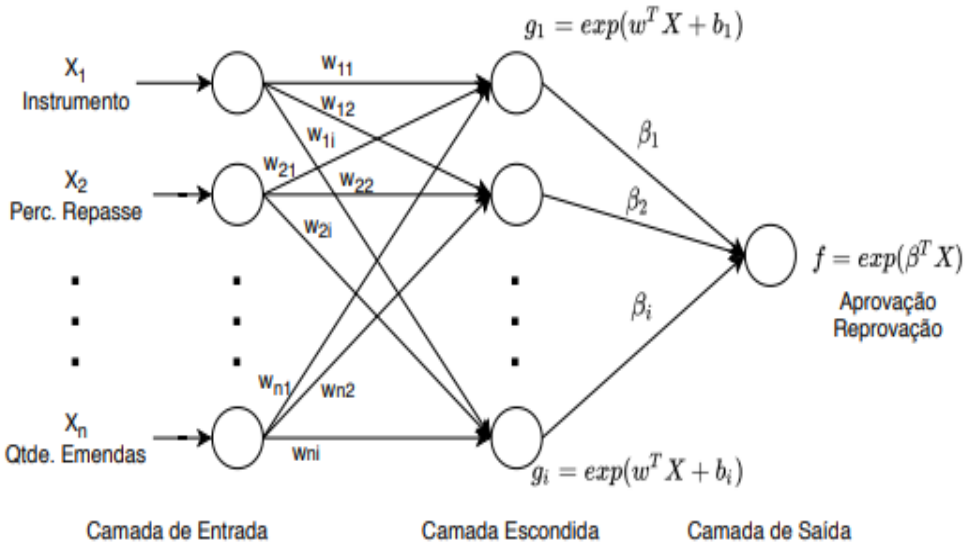
Após a implementação das equações acima, o modelo Xgbooster pode ser utilizado tanto como classificador como preditor, sendo eficiente com grandes volumes de dados.

### 2.2.3 *Multilayer perceptron* (MLP)

O Multilayer Perceptron (MLP) é uma forma clássica de rede neural artificial, conhecido por suas habilidades de mapeamento e frequentemente usado para aproximar funções arbitrárias (WIDIASARI; NUGROHO; WIDYAWAN, 2017). Os principais hiper parâmetros do DMLP são: taxa de aprendizado, número máximo de interações e função de regularização de perda (ORIMOLOYE et al., 2020; WIDIASARI; NUGROHO; WIDYAWAN, 2017).

A Figura 1 apresenta o algoritmo DMLP, sendo a camada mais à esquerda conhecida como camada de entrada, que consiste em um conjunto de neurônios representando as características de entrada ou atributos de entrada do modelo (*features*). Cada neurônio na camada oculta transforma os valores da camada anterior com uma soma linear ponderada, seguido por uma função de ativação não linear como a função hiperbólica. A camada de saída recebe os valores da última camada oculta e os transforma em valores de saída, ou seja, os *outputs* do modelo (MANJUNATH; MARIMUTHU; GHOSH, 2023; ORIMOLOYE et al., 2020; WIDIASARI; NUGROHO; WIDYAWAN, 2017).

Figura 1 – Funcionamento MLP




Fonte: Adaptado de (MANJUNATH; MARIMUTHU; GHOSH, 2023).

## 3. Metodologia

### 3.1 Amostra dos dados

Para desenvolver a presente pesquisa, foram selecionados ativos listados na Brasil Bolsa Balcão (B3), especificamente os que fizeram parte do IBOVESPA (IBOV) no primeiro trimestre de 2023, de modo estático, assim como (BARROSO; CARDOSO; MELO, 2021), sendo coletados por meio do *Website Yahoo Finance*. Em seguida, foram coletados dados de séries históricas dos preços (fechamento, abertura, máximo e mínimo), já ajustados por proventos, além da quantidade de negócios, volume financeiro e data (ano, mês e dia) das transações realizadas, dos códigos e ativos avaliados na granularidade diária. Os dados foram coletados considerando o período de janeiro de 2014 a janeiro de 2023, um período não muito curto e não muito longo, seguindo o mesmo padrão de tempo de alguns estudos similares (CHEN; HAO, 2018; WANG et al., 2020). O ativo livre de risco utilizado neste estudo foi a taxa Selic, que está disponível no site do Banco Central do Brasil (BACEN) com frequência diária, semelhante à metodologia utilizada por (Silva et al., 2024; Dimaziario Fernandes, 2020).

Como sugerem estudos anteriores, a combinação de diferentes indicadores técnicos melhora a qualidade da previsão dos modelos (THENMOZHI; SARATH CHAND, 2016; ZHOU et al., 2019) e, por isso, foram selecionados indicadores financeiros para servirem de entrada no modelo. A Tabela 1 apresenta todas as variáveis de entrada do modelo, que foram encontradas na literatura relacionada, e suas respectivas definições e referências.

Tabela 1- Indicadores financeiros

| N | Fórmula | Definição | Referência |
|---|---|---|---|
| 1 | $R_1 = \ln\left(\frac{fechamento_t}{fechamento_{t-1}}\right)$ | Retorno logarítmico calculado a partir dos preços do dia em relação ao preço do dia anterior (t-1) | CHEN et al., 2021; PAIVA et al., 2019) |
| 5 | $M = \frac{fec[h] + fec[h-1] + \ldots.}{5}$ | Mede a média móvel dos fechamentos dos | (MANJUNATH; MARIMUTHU; GHOSH, 2023) |

| | | | |
|---|---|---|---|
| | | últimos cinco dias de cotação | |
| 7 | $ME = ME_{[t-1]} + F * (P - ME_{[t-1]})$<br>Onde:<br>ME[t] = média exponencial<br>ME[t-1] = média exponencial anterior<br>F = Fator (F = 2 / (N+1)), onde N = número de dias para o cálculo da ME<br>P = preço atual | Média móvel exponencial, que dá um peso maior aos preços mais recentes dos ativos, considerando uma janela de 5 dias. | (BOONGASAME; SONGRAM, 2023) |
| 11 | Volume médio | Traz o volume médio de negociação dos ativos | |
| 16 | LS = M + d * DP<br>LI = M – d * DP<br>onde:<br>i = 0..n-1<br>M = média móvel simples<br>LS = limite superior<br>LI = limite inferior<br>DP = desvio padrão | Bollinger Bands é uma ferramenta de análise técnica que ajuda a identificar possíveis mudanças de tendência, além de mostrar a volatilidade do mercado. | (CHEN et al., 2018) |
| 17 | P(t) = P(t-1) + AF x (EP(t-1) – P(t-1)),<br>Onde:<br>P(t) = valor atual do indicador;<br>P(t-1) = valor do indicador um período antes;<br>AF = fator de aceleração, aumenta em 0.02 no intervalo [0.02;0.2]; | O indicador técnico parabólico, também conhecido como Parabolic SAR, é um indicador de seguimento de tendência que ajuda a identificar pontos de reversão de tendência em um gráfico de preço. A sigla SAR | (SON; PARK; HUH, 2019) |

| | | | |
|---|---|---|---|
| | EP(t-1) = valor mínimo/máximo do preço no período anterior. | significa "stop and reverse", indicando que o indicador é usado para ajudar os *traders* a identificar pontos de entrada e saída em uma posição de negociação. | |
| 18 | $\text{Medprice} = \frac{min. + max.}{2}$ | Mede o preço mediano em relação aos valores máximos e mínimos, no dia t. | (SAKHARE; SHAIK; SAHA, 2023) |
| 19 | OBV (On balance volume)<br>Se Fec[n] > Fec[n-1] then OBV[n] = OBV[n-1] + Vol$[n]<br>Se Fec[n] < Fec[n-1] then OBV[n] = OBV[n-1] – Vol$[n]<br>Se Fec[n] = Fec[n-1] then OBV[n] = OBV[n-1]<br>obs: Se parametrizado, por exemplo, em 100, o cálculo é refeito em cada data usando sempre apenas os últimos 100 pregões. | Analisa o volume de negociações | (BASAK et al., 2019) |
| 20 | $MAC = EMA_n - EMA_{ni}$ | Média móvel convergente e divergente | (BARROSO; CARDOSO; MELO, 2021; BASAK et al., 2019) |
| 21 | $IFR = \frac{100 * altas}{(altas + baixas)}$<br>onde:<br>altas [h] = média das últimas 9 altas<br>baixas [h] = média das últimas 9 baixas<br>alta [h] = fec [h] – fec [h – 1]<br>baixa [h] = fec [h – 1] – fec [h]<br>fec [h] = fechamento no dia | Índice de força relativa: mede a velocidade e a mudanças em preços. | (BASAK et al., 2019; ZHOU et al., 2019) |

| | | | |
|---|---|---|---|
| | obs: Altas e baixas não podem ser negativas. Se o resultado da fórmula for negativo, ele é truncado para o 0 (zero). | | |
| 22 | $$\%K = \frac{100 * (fec - min5)}{(max5 - min5)}$$ $$\%D = \frac{100 * (m3fec - m3min5)}{(m3max5 - m3min5)}$$ onde: max5 = cotação máxima dos últimos 5 dias min5 = cotação mínima dos últimos 5 dias m3fec = média dos 3 últimos fechamentos m3max5 = média dos 3 últimos max5 m3min5 = média dos 3 últimos min5 | Stochastic é um indicador de duas linhas que os *traders* podem utilizar em qualquer gráfico. Essas duas linhas são as linhas %K e %D, que se movem entre 0 e 100. Quando o indicador Stochastic está elevado, o preço do instrumento fecha perto do topo da amplitude de 14 períodos. | (MANJUNATH; MARIMUTHU; GHOSH, 2023) |
| 23 | $$Momentum = P - P[X]$$ Onde: P = preço de fechamento P[x] = preço de fechamento de x dias atrás | O Momentum é um indicador que mede a velocidade com que os preços variam quando comparados aos níveis dos valores atuais de um determinado ativo. | (MANJUNATH; MARIMUTHU; GHOSH, 2023) |

Fonte: Elaborado pelos autores a partir da literatura relacionada.

A Tabela 1 é dividida em uma coluna com a definição dos cálculos, uma coluna com a definição do indicador e uma coluna com um estudo ligado à área de seleção de portfólio de investimentos que implementou o indicador. A variável de saída do modelo será uma variável binária, assumindo o valor 1 quando o ativo atinge o valor alvo de ganho e 0, caso contrário,

semelhante a abordagem realizada por Silva et al. (2024). O valor alvo de ganho foi definido em uma coluna auxiliar que terá como base a comparação com o *benchmark* de mercado, sendo que o objetivo primário do modelo será obter ganhos substanciais ao *benchmark* de mercado. Assim, nesta pesquisa, será considerado o índice IBOV, seguindo a seguinte premissa:

i. Se a diferença percentual entre o retorno logarítmico do ativo e o retorno logarítmico do *benchmark* for maior ou igual a 2%, a variável *target* assume 1 para esse ativo, indicando que ele superou o desempenho do *benchmark*.
ii. Se a diferença percentual for menor que 2% (ou seja, o ativo subir menos que 2% em relação ao *benchmark* ou cair), a variável *target* para esse ativo assume 0, indicando que ele não atingiu o limiar de 2% de desempenho superior ao *benchmark*.

Ressalta-se que os atributos de entrada do modelo listados na Tabela 1 passaram pelo método de seleção de atributos (*features selection*), que tem por objetivo selecionar os atribudos de entrada mais realevantes ao modelo e será discutido na próxima seção.

### 3.2 Seleção de atributos

A qualidade dos dados de entrada no modelo influencia diretamente os resultados, sendo que a escolha dos atributos pode causar *overfitting* e baixa precisão do modelo (PRASTYO; ARDIYANTO; HIDAYAT, 2020). Diante desse contexto, selecionar os atributos que estão alinhados ao problema que se espera resolver é uma parte fundamental para melhorar o desempenho do modelo (FERRI et al., 1994). Dentre as possibilidades para selecionar atributos, destaca-se a redução de dimensionalidade quando há muitas variáveis (MANJUNATH; MARIMUTHU; GHOSH, 2023) e a seleção de atributos com base em regras e modelos específicos como uso de programação genética para selecionar indicadores técnicos (PIMENTA et al., 2018a).

Neste estudo, foi utilizada a técnica de seleção de atributos *Extratree*, um método de seleção de atributos em que árvores de decisão são combinadas de maneira aleatória para formar um conjunto de árvores, que são então utilizadas para estimar a importância dos atributos de acordo com um problema a ser solucionado (GEURTS; ERNST; WEHENKEL, 2006). A principal vantagem do modelo *Extratree* em relação a outros métodos de seleção de atributos é a sua capacidade de explorar o espaço de atributos de forma mais ampla e robusta, pois a combinação aleatória de árvores torna o modelo menos suscetível a problemas de *overfitting* ou *underfitting* (SHARMA et al., 2019).

### 3.5 Treinamento e avaliação do modelo

Para realizar o treinamento do modelo, foi adotado o método de validação cruzada. A validação cruzada é um procedimento que envolve a divisão do conjunto de dados em k partições. O modelo é treinado em k-1 partições e avaliado em uma partição de teste diferente em cada iteração. Esse processo é repetido k vezes, sempre com a partição de teste diferente (BENGIO; GRANDVALET, 2004). O resultado da validação cruzada é geralmente interpretado por meio da construção de intervalos de confiança para métricas como precisão ou acurácia do modelo. Neste estudo, adotamos k = 10 partições, seguindo a abordagem recomendada por referências anteriores (SONG et al., 2014; TENG; MA, 2022).

Após o treinamento do modelo pela validação cruzada, os modelos foram avaliados pela matriz de confusão, que exibe suas classes atuais em relação às classes previstas, ou seja, testa o nível de assertividade do modelo (FAVERO; BELIFIORE, 2017). A partir da matriz de confusão, foram calculadas métricas de desempenho do modelo, i) precisão, ii) *recall*, iii) acurácia e IV) F1 *score*. A Tabela 2 abaixo apresenta as definições:

Tabela 2- Métricas de avaliação dos modelos de *machine learning*

| Métrica | Descrição | Equação | Referência |
|---|---|---|---|
| Precisão | Precisão: Mede a proporção de verdadeiros positivos (amostras corretamente classificadas como positivas) em relação ao total de positivos previstos. | $\frac{TP}{TP + FP}$ | (HUANG, 2012b; KONG; YUN; KIM, 2023) |
| *Recall* | *Recall*: Também chamado de Sensibilidade, mede a proporção de verdadeiros positivos em relação ao total de positivos reais. | $\frac{TP}{TP + FN}$ | (HUANG, 2012b; KONG; YUN; KIM, 2023) |
| Acurácia | Acurácia: Mede a proporção de todas as previsões corretas (verdadeiros positivos e verdadeiros negativos) em relação ao tamanho total do conjunto de dados. | $\frac{TP + TN}{TP + FN + TN + FP}$ | (KONG; YUN; KIM, 2023) |
| F1 *score* | F1 *Score*: É uma métrica que combina precisão e *recall* em uma única pontuação. É útil quando se deseja um equilíbrio entre a precisão e a capacidade de recuperar todos os verdadeiros positivos. | $\frac{2 * \text{Precisão} + Recall}{Precisão + Recall}$ | (ZHOU et al., 2019) |

Fonte: Elaborado pelos autores a partir da literatura relacionada.

### 3.7 Modelo de otimização de portfólio

Dado um conjunto de ativos disponíveis para investimento, busca-se alocar recursos de forma ótima, utilizando um modelo de otimização de portfólios. O modelo utilizado trata de um modelo de otimização que tem como funções objetivos a minimização do risco, representada pela volatilidade, e a maximização do retorno do portfólio, semelhante ao estudo realizado por (PAIVA

et al., 2019; PIMENTA et al., 2018b). O modelo teve a inclusão do coeficiente de aversão ao risco $\lambda$ , que foi introduzido por (CHANG; YANG; CHANG, 2009) para equilibrar o *trade off* entre risco e retorno. O portfólio faz a alocação de recursos a partir de um portfólio inicial, permitindo assim o rebalanceamento diário, a fim de realizar o acompanhando do retorno do portfólio ao longo do tempo. Além disso, incorporou-se uma restrição de alocação máxima de 15%, conforme (RUBESAM; BELTRAME, 2013), que limita o valor a ser investido em cada ativo do portfólio com o intuito de evitar a concentração de investimento em ativos especificos.

Abaixo segue modelo de otimização:

$$Max.E = \sum_{i=1}^{N} X.\bar{u} \tag{11}$$

$$Min.V = \lambda \left| \sum_{i=1}^{N} \sum_{J=1}^{N} X_I \ .X_J.\sigma_{ij} \right| \tag{12}$$

$$\textbf{Sujeito} \begin{cases} \sum_{i=1}^{N} \boldsymbol{X_i = 1} \\ \boldsymbol{0 \leq X_i \leq 0.15} \end{cases} \tag{13}$$

Onde:

E = retorno médio esperado da carteira;

V= variância da carteira;

ū = média de retorno por ativo;

$X_i$ = proporção do ativo *i* no portfólio;

$\sigma_{ij}$ = covariância;

N= número de ativos do portfólio;

k= custo de transação proporcional a cada transação;

$V_{it}$=valor de abertura do ativo *i* no tempo *t*;

F=custo de transação fixo, proporcional à quantidade de operações

O modelo de otimização será implementado ao longo de um período de três anos, considerando tanto o grupo de ativos selecionados por cada modelo de *machine learning* quanto todos os ativos da amostra coletada do Ibovespa. O objetivo é comparar o desempenho do modelo de otimização com a pré-seleção de ativos em relação ao modelo sem essa pré-seleção. Especificamente, a validação foi realizada para o período de 01/01/2020 a 01/01/2023, com o portfólio sendo rebalanceado diariamente.

### 3.7.1 Custos de transação

O modelo de otimização foi analisado inicialmente sem considerar os custos de transação e, posteriormente, os custos de transação foram incorporados para avaliar seu impacto no desempenho do modelo, de modo semelhante à metodologia de (PAIVA et al., 2019; WANG et al., 2020). Essa abordagem permitirá uma compreensão mais profunda da influência dos custos de transação na eficácia do modelo de otimização. A restrição de custo de transação baseia-se no estudo de Valle; Meade e Beasley (2014), sendo testados os custos de 0.10 bps, 0.50 bps e 1 bps, assim como em outros estudos (GUERARD; MARKOWITZ; XU, 2015; MA; HAN; WANG, 2021; PAIVA et al., 2019). Abaixo segue a equação para calcular os custos de transação:

$$\begin{cases} \boldsymbol{G_i \geq k.X_i.V_{it} + F, i = 1, \ldots . N} \\ \boldsymbol{0 \leq X_i . \forall i = 1, \ldots , N} \end{cases} \quad (\mathbf{14})$$

$G_i$=custo de transação de cada ativo;

k= custo de transação proporcional a cada transação;

$V_{it}$=valor de abertura do ativo i no tempo t;

### 3.7.2 Algoritmo de otimização

Para efetuar a otimização do portfólio de investimento, foi adotado o algoritmo *Sequential Least-Squares Quadratic Programming* (SLSQP) por meio da biblioteca *SciPy* na linguagem *Python*. O algoritimo SLSQP foi proposto por (KRAFT, 1998) com a finalidade de resolver problemas de otimização não linear com restrições. A Tabela 3 apresenta o passo a passo do algoritmo.

Tabela 3- Esquema do algoritmo SLQSP

Algoritmo: Esquema do algoritmo SLQSP

Inicialização do algoritmo:

Determine os parâmetros como função objetivos e restrições.

Forneça valores iniciais para função que deseja otimizar.

Construção da função quadrática:

Construa um modelo quadrático aproximado da função objetivo com base nos valores iniciais dos parâmetros.

**while** enquanto o critério de parada não for atingido

Realize otimização quadrática local

Leve em consideração as restrições durante a otimização.

Atualize os parâmetros com base no novo ponto de encontro.

Verifique se a convergência foi alcançada.

Se a convergência não for alcançada retorne a etapa de construção do modelo.

**end**

Fonte: Elaborado pelos autores a partir de (KRAFT, 1998).

### 3.8 Processo de simulação (*backtesting*)

Nesta seção, é apresentado o processo de simulação (*backtesting*) utilizado para medir o desempenho das carteiras no período de janeiro de 2020 a janeiro de 2023. As carteiras terão rebalanceamento das alocações realizadas diariamente, a fim de acompanhar o desempenho do retorno ao longo do período de *backtesting*.

Para cada portfólio formado a partir de *machine learning*, foram empregadas as seguintes medidas de comparação do desempenho de carteiras, conforme estudos anteriores (GUIMARÃES JÚNIOR; CARMONA; GUIMARÃES, 2015; RUBESAM; BELTRAME, 2013): média de excesso de retorno, desvio-padrão dos retornos da carteira, Índice de Sharpe (IS), Índice de Treynor, Alfa de Jensen e VAR (valor e risco).

O Índice de Sharpe representa a relação prêmio pelo risco e o risco do ativo, de acordo com Sharpe (1964), e é calculado conforme a equação (15).

$$IS = \frac{(R_p - R_f)}{Dp} \qquad (15)$$

Onde $R_p$ é o retorno do portfólio, $R_f$ é o ativo livre de risco e Dp é o risco do portfólio.

Neste estudo, o ativo livre de risco foi a taxa Selic, que foi coletada diariamente do site do banco central, conforme foi utilizada no estudo de Dimarzio; Matias Filho e Fernandes (2020).

Outro indicador utilizado foi o índice de Treynor. O Índice de Treynor representa o prêmio pelo risco gerado por uma carteira por unidade do seu risco não diversificável (GUIMARÃES JÚNIOR; CARMONA; GUIMARÃES, 2015). A equação (16) apresenta o cálculo do Índice de Treynor.

$$IT = \frac{(R_p - R_f)}{\beta p} \tag{16}$$

Onde o $\beta p$ representa o coeficiente beta, que é uma medida da sensibilidade do retorno do portfólio em relação ao retorno da carteira considerada de mercado.

O beta é usado para quantificar o risco sistemático, ou seja, o risco que não pode ser eliminado através da diversificação do portfólio (HÜBNER, 2005). O beta é calculado pela equação (17), sendo que $R_m$ representa o retorno de mercado, que nesse caso serão médias de retorno do IBOV.

.

$$\beta p = \frac{Cov(R_p, R_m)}{var(R_m)} \tag{17}$$

O Alfa de Jensen, também conhecido como excesso de retorno, é uma medida de desempenho que indica a capacidade de um investimento ou carteira de superar ou ficar aquém do retorno esperado com base no risco sistemático assumido (JENSEN, 1968). O Alfa de Jensen é calculado pela equação (18) :

$$\alpha_i = (R_p - R_f) - \beta p\big(R_m - R_f)\big] \tag{18}$$

Onde:

αp= alfa de Jensen do portfólio;

$R_f$ = retorno do ativo livre de risco;

$\beta p$ = beta da carteira;

$R_m$ = retorno médio do mercado.

O VAR (*Value at Risk*) é uma medida estatística utilizada para estimar a perda máxima esperada de um investimento ou carteira em um determinado intervalo de confiança(WU et al.,

2023). O VAR é calculado a partir da equação (1*), sendo que o presente estudo utilizou um nível de confiança de 95 %, sendo a variável Z correspondente ao número de desvios abaixo ou acima da área delimitada pela curva normal.

$$VAR = R_p - \left(Z.D_p.Q\right) \quad (19)$$

### 3.9 Software de análise de dados e bibliotecas utilizadas na dissertação

As análises dos dados do artigo foram desenvolvidas utilizando a linguagem Python versão 3.11.1, que se trata de uma linguagem de programação *open source* amplamente utilizada para análise de dados e *machine learning*, tanto no ambiente corporativo como em pesquisas acadêmicas. Foram ultilizadas as seguintes bibliotecas na presente pesquisa:

- Numpy: biblioteca para computação científica que tem funções da área de matemática, álgebra linear, geração de número randômicos (PEREZ et al., 2022a);
- Pandas: biblioteca para manipulação e estruturação da base de dados que possui funções para analisar e manipular dados (PATEL M. K., 2020);
- Matplotlib: biblioteca para plotagem de gráficos (HUNTER J et al., 2017);
- Seaborn: biblioteca de visualização de dados estatísticos que integra diretamente com o Matplotlib e com funções do Pandas (Waskom, 2021);
- Scipy : biblioteca que possui funções da área de matemática, ciências e engenharia (PEREZ et al., 2022);
- Scikit-Learn: biblioteca para uso de técnicas de inteligência artificial, *machine learning* e *deep learning* (HACKELING G, 2017);
- Statsmodels: biblioteca para análise estatística e econométrica (Seabold and Perktold, 2010);
- Gurob: biblioteca que resolve problemas de programação linear, quadrática, linear inteira, mista e quadrática inteira mista (Santos and Toffolo, 2020);
- Keras: biblioteca que resolve para uso de técnicas de *deep learning* (Moolayil, 2019).

## 4. Resultados e discussões

### 4.1 Pré-processamento dos dados

O primeiro passo após a coleta de dados e o cálculo dos indicadores financeiros foi realizar o tratamento dos dados, removendo valores ausentes e *outliers*, conforme recomendado por (PIMENTA et al., 2018). Para realizar as análises, foi selecionado um grupo de 62 ativos que possuem informações completas ao longo do período de análise. A Tabela 4 exibe os detalhes dos ativos escolhidos para análise como opções de investimento ao longo deste estudo.

Tabela 4- Relação de ativos utilizados na pesquisa e seus respectivos setores

| Abreviatura | Setor | Abreviatura | Setor | Abreviatura | Setor |
|---|---|---|---|---|---|
| ALPA4 | Consumer Cyclical | COGN3 | Consumer Defensive | GOLL4 | Industrials |
| ABEV3 | Consumer Defensive | CPLE6 | Utilities | HYPE3 | Healthcare |
| AMER3 | Consumer Cyclical | CSAN3 | Energy | ITSA4 | Industrials |
| ARZZ3 | Consumer Cyclical | CPFE3 | Utilities | ITUB4 | Financial Services |
| B3SA3 | Financial Services | CYRE3 | Consumer Cyclical | JBSS3 | Consumer Defensive |
| BPAN4 | Financial Services | DXCO3 | Basic Materials | RENT3 | Industrials |
| BBDC3 | Financial Services | ECOR3 | Industrials | LREN3 | Consumer Cyclical |
| BBDC4 | Financial Services | ELET3 | Utilities | MGLU3 | Consumer Cyclical |
| BRAP4 | Financial Services | ELET6 | Utilities | MRFG3 | Consumer Defensive |
| BBAS3 | Financial Services | EMBR3 | Industrials | BEEF3 | Consumer Defensive |
| BRKM5 | Basic Materials | ENEV3 | Utilities | MRVE3 | Consumer Cyclical |
| BRFS3 | Consumer Defensive | EGIE3 | Utilities | MULT3 | Real Estate |
| CCRO3 | Industrials | EQTL3 | Utilities | PETR3 | Energy |
| CMIG4 | Utilities | EZTC3 | Real Estate | PETR4 | Energy |
| CIEL3 | Technology | FLRY3 | Healthcare | PRIO3 | Energy |
| GOAU4 | Basic Materials | GGBR4 | Basic Materials | YDUQ3 | Consumer Defensive |
| QUAL3 | Healthcare | CSNA3 | Basic Materials | TOTS3 | Technology |
| RADL3 | Healthcare | SLCE3 | Consumer Defensive | UGPA3 | Energy |
| SBSP3 | Utilities | TAEE1 | Utilities | USIM5 | Basic Materials |
| SANB1 | Financial Services | VIVT3 | Communication Services | VALE3 | Basic Materials |
| SMTO3 | Basic Materials | TIMS3 | Communication Services | WEGE3 | Industrials |

Fonte: Elaborado pelos autores com os dados da pesquisa a partir do yfinance.

Os ativos considerados para análise exibem uma diversificação significativa em termos de setores, pois o Ibovespa é considerado o *benchmark* mais representativo e diversificado da bolsa de valores brasileira (BARROSO; CARDOSO; MELO, 2021).

### 4.2 *Feature selection*

A partir dos ativos listados na Tabela 4 e das variáveis relacionadas na Tabela 1, aplicou-se o processo de seleção de características utilizando o modelo *ExtraTreeClassifier*. O objetivo desse processo foi identificar quais variáveis são as mais relevantes para resolver o problema em questão. A Tabela 5 apresenta o resultado da implementação do modelo *ExtraTreeClassifier* para seleção de características.

Tabela 5- Ordenação da importância pelo modelo ExtraTreeClassifier

| Feature | Importance |
|---|---|
| Retorno Logarítmico | 0.267555339 |
| Stoch_K | 0.108488935 |
| Stoch_D | 0.066682119 |
| RSI | 0.04110329 |

| | |
|---|---|
| Open | 0.040829109 |
| Volume | 0.040403783 |
| MACD | 0.03953499 |
| Fechamento | 0.039386798 |
| MACD_Signal | 0.038473856 |
| Preço ajustado | 0.037606607 |
| Alta | 0.036830336 |
| baixa | 0.036535361 |
| PSAR | 0.03610711 |
| BB_Lower | 0.035593021 |
| BB_Upper | 0.035163586 |
| SMA | 0.033594497 |
| OBV | 0.033347527 |
| EMA | 0.032763737 |

Fonte:Elaborado pelos autores.

A partir da análise da Tabela 5, observa-se que o retorno logarítmico dos ativos é a característica mais relevante seguindo dos indicadores Stochastic K e D, que avaliam a tendência de mercado, juntamente com o indicador RSI. A partir da seleção de características pelo método *ExtraTreeClassifier*, seguiu-se com apenas os dez indicadores mais relevantes para tentar dar sinais de quais ativos deve-se investir, assim como no estudo de (BHUVANA CHANDRIKA; VIJAYANAND; RAO, 2021), que ultilizou essa mesma metodologia para elencar as dez melhores alternativas.

### 4.3 Seleção de investimentos por meio dos modelos de *machine learning*

Para realizar a seleção de ativos que irão compor os portfólios, foram ultilizados os métodos de *machine learning*: *Random forest*, *XGBboost e Multilayer Perceptron*. Os modelos foram treinados e testados considerando as métricas *recalls*, f1 *score*, precisão e acurácia, conforme descrito na Tabela 2. A Tabela 6 apresenta o resultado da avaliação de cada modelo.

Tabela 6- Métricas de avaliação dos modelos de *machine learning*

| **Modelo** | ***Precision*** | ***Accuracy*** | **F1-*Score*** | ***Recall*** |
|---|---|---|---|---|
| XGBoost | 0.760026 | 0.911634 | 0.63207 | 0.540991 |
| Multilayer Perceptron | 0.774983 | 0.912242 | 0.627904 | 0.527746 |
| Random Forest | 0.757986 | 0.910929 | 0.628243 | 0.536424 |

Fonte:Elaborado pelos autores a partir dos dados coletados.

Na Tabela 6, são exibidos os resultados de cada modelo para diferentes métricas de avaliação, enfocando a capacidade de cada modelo em prever as classes positivas. Observa-se que os desempenhos dos modelos foram bastante similares, sendo que, em termos de acurácia, todos apresentaram resultados próximos. No entanto, o modelo *Multilayer Perceptron* se destacou com

um desempenho um pouco superior na métrica de precisão, atingindo aproximadamente 77,49%. Por outro lado, o modelo Xgboost liderou na métrica de *recall*, com cerca de 54%, demonstrando sua habilidade em identificar corretamente as instâncias positivas. Além disso, o modelo Xgboost também obteve o melhor resultado na métrica F1 *score*, alcançando cerca de 63,20%. Isso sugere que o modelo Xgboost seja uma escolha mais eficaz para esse problema de previsão.

Comparando os resultados com o estudo de Silva et al. (2024) e Paiva et al. (2019), que utilizaram o método SVM para classificação e foram aplicados no Brasil, a abordagem proposta em todos os modelos obteve uma precisão e acurácia superiores. Silva et al. (2024) alcançaram uma precisão de 56 % e acurácia de 61%, enquanto Paiva et al. (2019, p.13) encontraram uma precisão média de 54,97%.

A partir dos resultados obtidos com a previsão de cada modelo, foram determinados grupos de investimentos que constituíram os portfólios recomendados por cada modelo. A Tabela 7 destaca os ativos selecionados por cada modelo, que serão utilizados no processo de otimização do portfólio de investimento.

Tabela 7- Ativos selecionados por cada modelo

| **Nº ativos** | **XGBoost** | ***Multilayer Perceptron*** | ***Random Forest*** |
|---|---|---|---|
| 1 | MGLU3 | MGLU3 | PRIO3 |
| 2 | PRIO3 | PRIO3 | USIM5 |
| 3 | AMER3 | GOLL4 | GOLL4 |
| 4 | GOLL4 | AMER3 | MGLU3 |
| 5 | USIM5 | USIM5 | BPAN4 |
| 6 | ELET3 | COGN3 | AMER3 |
| 7 | COGN3 | BPAN4 | ELET3 |
| 8 | CSNA3 | ELET3 | GGBR4 |
| 9 | BPAN4 | BRKM5 | CIEL3 |
| 10 | BRAP4 | ENEV3 | CSNA3 |
| 11 | PETR3 | CSNA3 | BRKM5 |
| 12 | GGBR4 | RENT3 | GOAU4 |
| 13 | DXCO3 | MRFG3 | BRFS3 |
| 14 | MRVE3 | EMBR3 | EMBR3 |
| 15 | MRFG3 | ARZZ3 | SLCE3 |
| 16 | ENEV3 | PETR3 | PETR3 |
| 17 | CIEL3 | BRFS3 | ENEV3 |
| 18 | BRFS3 | ELET6 | MRFG3 |
| 19 | BRKM5 | MRVE3 | MRVE3 |
| 20 | CYRE3 | EZTC3 | |
| 21 | PETR4 | YDUQ3 | |
| 22 | VALE3 | CIEL3 | |

Fonte: Elaborado pelos autores.

### 4.4 Otimização de portfólio de investimento sem custos de transação

A Tabela 8 apresenta os resultados da análise das métricas de desempenho de modelos para os portfólios selecionados por cada modelo de *machine learning*, para todos os ativos pertencentes ao Ibovespa, para as carteiras ingênuas, que representam os pesos divididos por igual em cada ativo, e também o *benchmark* de mercado, que é o Ibovespa. Foram considerados os seguintes indicadores: o prêmio pelo risco, a média de retorno anualizado, o risco anualizado da carteira, a expectativa de retorno para cada portfólio, índice de Sharpe, índice de Treynor, Beta, Alfa de Jensen e o Var. Abaixo segue a Tabela 8.

Tabela 8 – Métricas de avalição dos modelos de *Machine Learning* em comparação com a carteira de mercado

| Indicador | Prêmio pelo risco | Índice de Sharpe | Retorno Anual | Risco Anual | Beta | Índice de Treynor | Alpha de Jensen | Var 95% | nº ativos |
|---|---|---|---|---|---|---|---|---|---|
| Portfólio MOO Xgboost | 7,76% | 0,24 | 14,42% | 0,32 | 0,90 | 0,09 | 0,10 | 2,68% | 22 |
| Portfólio MOO MLP | 8,21% | 0,25 | 14,87% | 0,33 | 0,93 | 0,09 | 0,11 | 2,87% | 22 |
| Portfólio MOO Rf | 12,45% | 0,39 | 19,11% | 0,32 | 0,89 | 0,14 | 0,15 | 2,71% | 19 |
| Portfólio MOO todos os ativos | 16,58% | 0,80 | 23,24% | 0,21 | 0,54 | 0,31 | 0,18 | 1,78% | 62 |
| Ibovespa | - | -0,29 | -2,53% | 0,31 | - | - | - | 2,58% | |
| Portfólio 1/n mlp | -19,80% | -0,48 | -13,14% | 0,42 | 1,25 | -0,16 | -0,17 | 3,50% | 62 |
| Portfólio 1/n Rf | -13,15% | -0,32 | -6,49% | 0,41 | 1,22 | -0,11 | -0,10 | 3,35% | 62 |
| Portfólio 1/n todos os ativos | -13,81% | -0,41 | -7,15% | 0,34 | 1,05 | -0,13 | -0,11 | 2,83% | 62 |
| Portfólio 1/n Xgboost | -17,38% | -0,43 | -10,72% | 0,41 | 1,24 | -0,14 | -0,14 | 3,35% | 62 |

Fonte:Elaborado pelos autores.

Ao analisar os indicadores da Tabela 8, começando pelo retorno anual, verifica-se que todos os portfólios superaram o *benchmark* de mercado, com exceção dos portfólios com a carteira ingênua. Percebe-se que o portfólio com todos os ativos teve um resultado superior aos demais

portfólios cerca de 23%, porém, vale ressaltar que o portfólio com todos os ativos possui um número maior de ativos, sendo que pode ser impraticável para um investidor gerenciar um portfólio com 62 ativos. No que tange os portfólios formados por *machine learning*, o portfólio *Random forest* apresentou um retorno anual de cerca de 19,11% ao ano, o que vai de encontro com a pesquisa de (MA; HAN; WANG, 2021), em que mesmo sendo aplicada em outro mercado e com outras variáveis de entrada o modelo *Random forest* apresentou melhor resultado. Em relação ao risco, observa-se que os portfólios mantiveram níveis próximos ao *benchmark* de mercado.

No entanto, ao avaliarmos o índice de Sharpe, fica claro que todos os portfólios superaram o índice de Sharpe do *benchmark*, indicando que tanto o portfólio *Random forest* quanto os outros demonstram uma relação risco-retorno mais favorável em comparação com o *benchmark*. Isso sugere que os portfólios apresentaram um desempenho superior em termos de equilíbrio entre risco e retorno em comparação com a carteira de mercado de referência.

O próximo indicador analisado foi o coeficiente beta de cada portfólio, sendo que todos indicaram uma menor volatilidade em relação ao mercado. Esses valores de beta confirmam que todos os portfólios têm uma tendência menor de seguir as flutuações do mercado em comparação com o *benchmark*, refletindo um menor risco sistemático em seus retornos.

Outro indicador relevante analisado é o índice de Treynor. No caso, o Portfólio Xgboost e MLP apresentaram um índice de 0,09, o que significa que está gerando cerca de 9% a mais de retorno do que um ativo livre de risco, levando em consideração seu nível de risco sistemático. O portfólio *Random forest* atingiu um índice de 0,14, o que significa que pode gerar 14% a mais de retorno do que um ativo livre de risco, considerando seu nível de risco não diversificável. Ambos os resultados são satisfatórios e sugerem que os três portfólios estão superando um investimento livre de risco em termos de retorno ajustado ao risco de mercado. O Portfolio *Random forest* apresentou o melhor desempenho nesse aspecto, seguido pelo MLP e Xgboost, mas todos demonstraram a capacidade de gerar retornos adicionais significativos em relação a um ativo livre de risco, considerando seus respectivos perfis de risco sistemático.

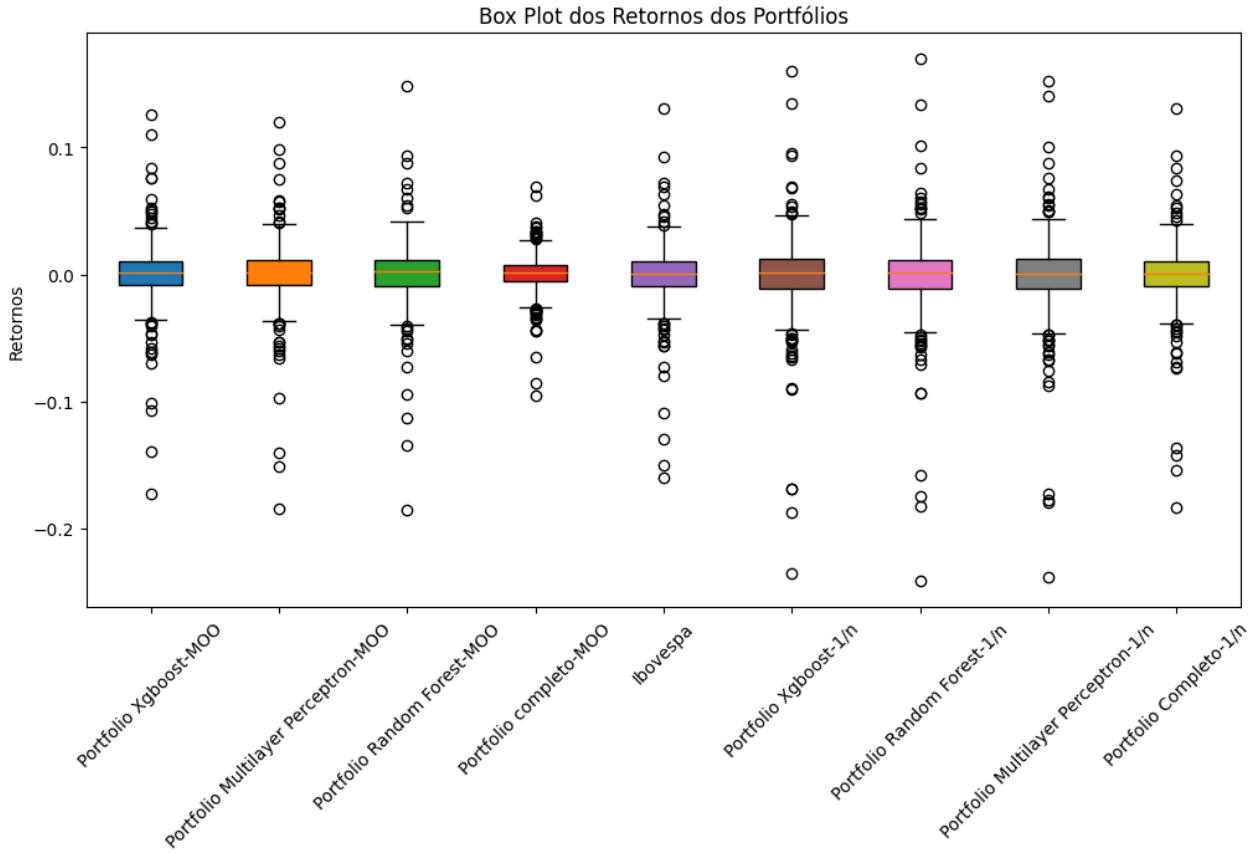
O próximo indicador avaliado foi o Alfa de Jensen. É importante notar que todos os portfólios otimizados apresentaram retornos superiores aos do mercado, refletindo Alfas de Jensen positivos. Mais uma vez, o Portfólio *Random forest*, juntamente com o portfólio com todos os ativos, demonstrou ser mais eficiente do que os demais, ambos com um Alfa de 0,15. Em seguida, temos os portfólios MLP com Alfa de Jensen 0.11 e o Xgboost com Alfa 0.10. Ressalta-se que os portfólios formados pela carteira ingênua obtiveram Alfa de Jensen negativo.

O último indicador analisado é o Var (*Value at Risk*), que mensura o risco de perdas em um portfólio em um determinado intervalo de confiança. No caso, o portfólio com todos os ativos

apresentou um Var de 1.78%,o Portfólio Xgboost apresentou um Var de 2.68%, o Portfólio MLP registrou 2,87% e o Portfólio Rf apresentou 2,71%. Esses valores indicam que, dentro do intervalo de confiança de 95 %, esses portfólios estão expostos a uma perda máxima de 1,78%, 2,68%, 2,87% e 2,71%, respectivamente. É relevante notar que todos os portfólios têm um Var próximo, sugerindo níveis de perdas esperadas semelhantes. Também podemos observar na Tabela 8 o Var do *benchmark* de mercado, que também esteve próximo ao resultado dos portfólios, cerca de 2,58%. Portanto, apesar de um desempenho positivo em outros indicadores, os portfólios selecionados via *machine learning* apresentaram um nível de risco um pouco superior em comparação com o mercado de referência, conforme indicado pelo Var.

Para aprofundar na análise dos resultados, é possível explorar a Figura 2, que apresenta um *box plot* mostrando a dispersão dos retornos ao longo do tempo. Essa representação gráfica permite uma visualização mais detalhada da variabilidade dos retornos dos portfólios.

Figura 2- *Box plot* retorno ao longo do tempo



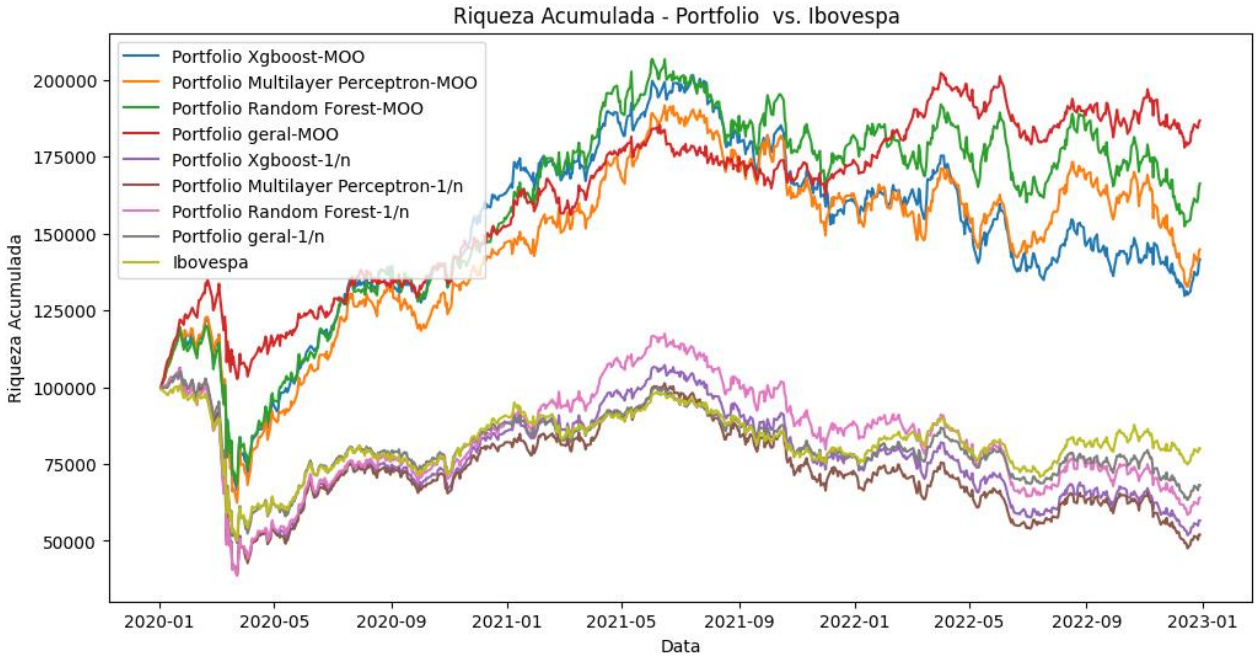

Fonte: Elaborado pelo autor.

A partir da análise da Figura 2, observa-se um *box plot* dos retornos das carteiras, onde a linha laranja representa a mediana dos retornos e os pontos acima e abaixo indicam a dispersão dos retornos, incluindo valores máximos, mínimos e *outliers*. Nota-se que tanto o Ibovespa quanto o

portfólio formado pela carteira ingênua apresentam muitos retornos no lado negativo, com poucos *outliers* positivos.

Outra análise importante diz respeito à riqueza acumulada ao longo do tempo, na qual os retornos de cada portfólio foram multiplicados por um investimento inicial de R$100.000,00. A Figura 3, apresentada abaixo, exibe o gráfico que ilustra a evolução da riqueza ao longo do período.

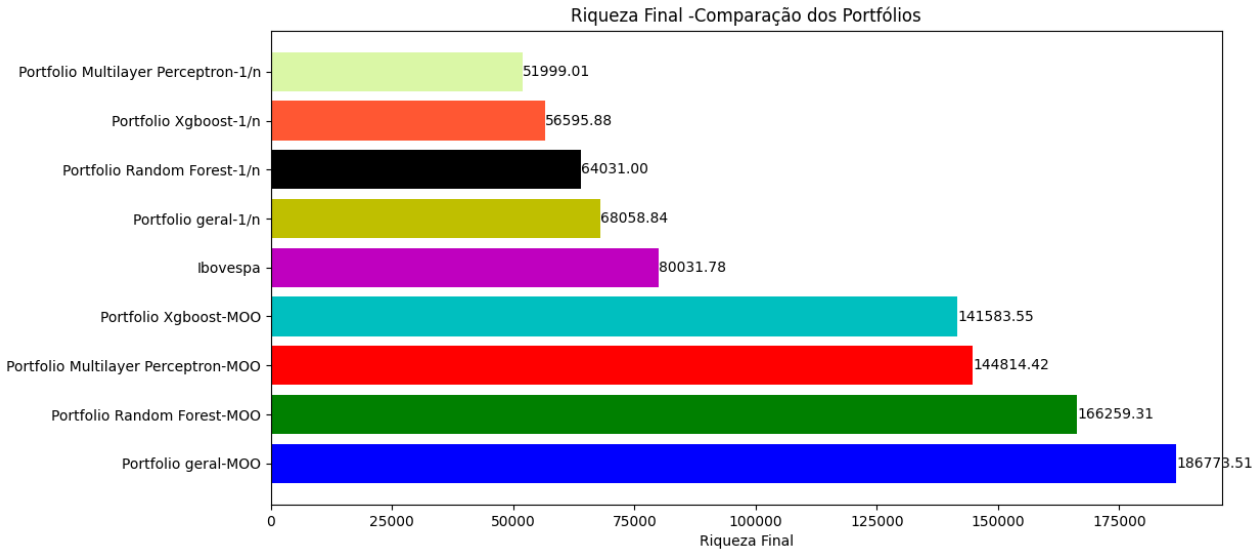
Figura 3- Comparação do excesso de retorno real acumulado ao longo do tempo entre as carteiras analisadas




Fonte: Elaborado pelo autor com os dados da pesquisa.

A partir da análise da Figura 3, fica evidente que ambos os portfólios otimizados apresentaram um desempenho significativamente superior ao *benchmark* de mercado e aos portfólios formados de maneira ingênua, superando-o ao longo de todo o período de análise. Além disso, é notável que os resultados dos portfólios foram bastante próximos, com o Portfólio *Random forest* e o portfólios de todos os ativos obtendo um desempenho particularmente maior no ano de 2022, o que contribuiu para que superasse os outros portfólios em termos de riqueza acumulada. Esse destaque no ano de 2022 refletiu-se no desempenho geral do Portfólio *Random forest* e do portfólio com todos os ativos em relação aos demais portfólios. A partir dessa análise, elaborou-se a Figura 4 com o valor da riqueza acumulada.

Figura 4- Riqueza acumulada ao longo do tempo




Fonte :Elaborado pelo autor.

Observa-se que o portfólio otimizado com todos os ativos obteve melhor desempenho, seguido dos portfólios otimizados com os métodos de *machine learning*: *Random forest*, MLP e *Xgboost*. Porém, ressalta-se que os modelos formados de *machine learning* apresentam um número reduzido de ativos, o que facilitaria a um investidor comum gerenciar os investimentos. Observa-se que o Ibovespa e os portfólios compostos pela estratégia ingênua obtiveram resultados negativos, ou seja, a riqueza incial ficou menor que a riqueza final após três anos.

### 4.4 Otimização de portfólio de investimento com custos de transação

A análise de custos de transação envolve a consideração de três cenários distintos: i) 0,10 (bps); ii) 0,50 (bps); e iii) 1,0 (bps). Para facilitar essa análise, elaboramos a Tabela 9, na qual apresentamos os resultados da avaliação dos indicadores de desempenho financeiro para cada um desses cenários.

Tabela 9 – Análise de desempenho dos portfolios

| Indicador | Portfólio MOO Xgboost | Portfólio MOO MLP | Portfólio MOO Rf | Portfólio MOO todos os ativos | Custo de transação |
|---|---|---|---|---|---|
| Prêmio pelo risco | 7,76% | 8,21% | 12,45% | 16,58% | 0.10 bps |
| Índice de Sharpe | 0,24 | 0,25 | 0,39 | 0,80 | 0.10 bps |
| Retorno Anual | 14,42% | 14,87% | 19,11% | 23,24% | 0.10 bps |
| Risco Anual | 0,32 | 0,33 | 0,32 | 0,21 | 0.10 bps |
| Beta | 0,90 | 0,93 | 0,89 | 0,54 | 0.10 bps |
| Índice de Treynor | 0,09 | 0,09 | 0,14 | 0,31 | 0.10 bps |
| Alpha de Jensen | 0,10 | 0,11 | 0,15 | 0,18 | 0.10 bps |
| Var 95% | 2,68% | 2,87% | 2,71% | 1,78% | 0.10 bps |
| Prêmio pelo risco | 4,995% | 6,323% | 9,994% | 11,083% | 0.50 bps |
| Índice de Sharpe | 0,15518281 | 0,18919226 | 0,306210897 | 0,529029139 | 0.50 bps |
| Retorno Anual | 10,995% | 12,323% | 15,994% | 17,083% | 0.50 bps |
| Risco Anual | 32,185% | 33,421% | 32,637% | 20,950% | 0.50 bps |

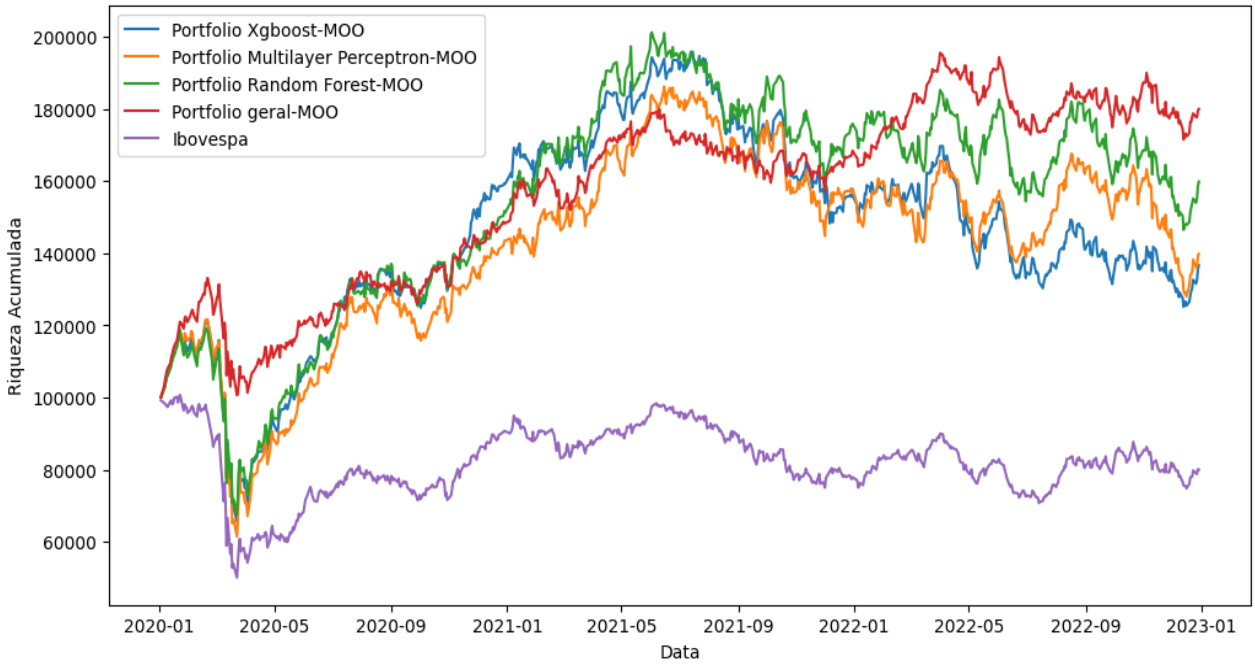
| | | | | | |
|---|---|---|---|---|---|
| Beta | 0,910177057 | 0,94058386 | 0,892388025 | 0,543556564 | 0.50 bps |
| Índice de Treynor | 0,054874318 | 0,06722369 | 0,111989512 | 0,203899819 | 0.50 bps |
| Alpha de Jensen | 7,30% | 8,70% | 12,25% | 12,46% | 0.50 bps |
| Var 95% | 2,72% | 2,91% | 2,73% | 1,78% | 0.50 bps |
| Prêmio pelo risco | -1,01% | 0,43% | 3,39% | 4,82% | 1 bps |
| Índice de Sharpe | -0,031055634 | 0,01271014 | 0,103479008 | 0,228573522 | 1 bps |
| Retorno Anual | 4,995% | 6,427% | 9,393% | 10,821% | 1 bps |
| Risco Anual | 32,4% | 33,6% | 32,8% | 21,1% | 1 bps |
| Beta | 0,916391632 | 0,94660872 | 0,898432795 | 0,549846324 | 1 bps |
| Índice de Treynor | -0,010967027 | 0,00450887 | 0,037767773 | 0,087680075 | 1 bps |
| Alpha de Jensen | 1,32% | 2,82% | 5,67% | 6,21% | 1 bps |
| Var 95% | 2,76% | 2,96% | 2,75% | 1,88% | 1 bps |

Fonte :Elaborado pelo autor.

A partir das análises realizadas, observa-se que os portfólios mantiveram um padrão consistente de desempenho, no qual o aumento nos custos fixos de transação resultou em uma diminuição do retorno médio, bem como de outras métricas como o Alfa de Jensen e o Índice de Sharpe. Observa-se também que no cenário onde os custos de transação são de 1 bps, o prêmio pelo risco é negativo, o que está de acordo com a análise realizada por Paiva et al. (2019). Também se observa que, mesmo quando comparados com o benchmark de mercado, os portfólios ainda apresentaram resultados superiores. Além disso, é importante ressaltar que todos os portfólios, nos três cenários considerados, obtiveram retornos positivos.

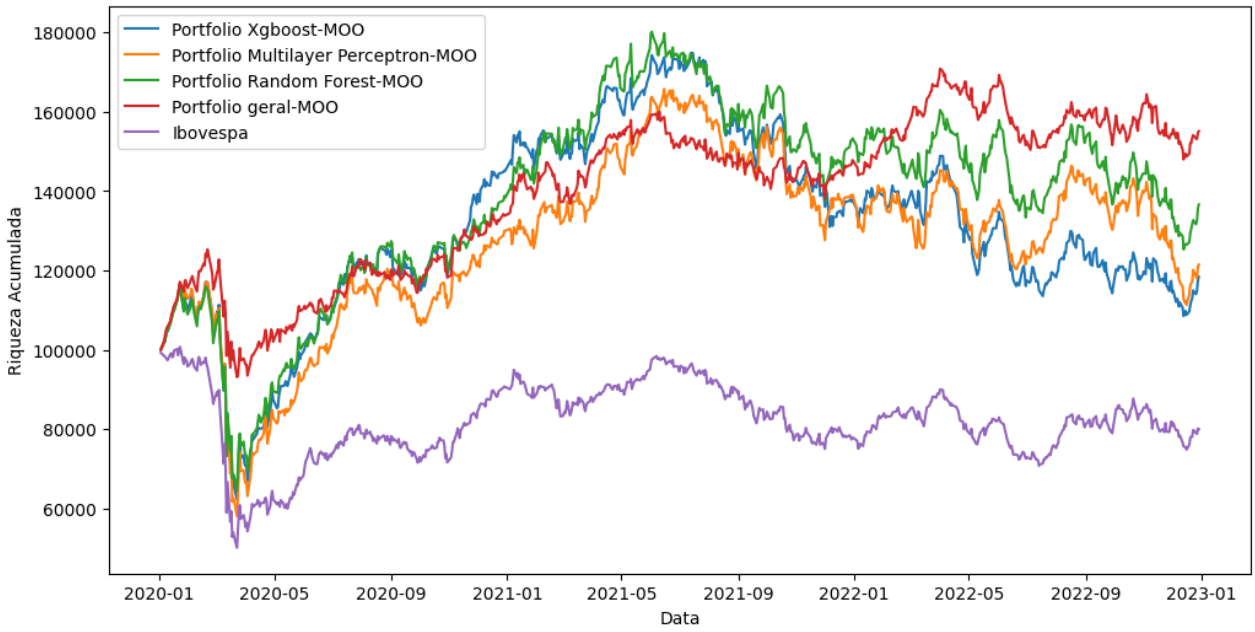
Com o objetivo de analisar o impacto dos custos de transação na acumulação de riqueza ao longo do tempo, foram criados os gráficos a seguir: Figura 5, Figura 6 e Figura 7. Cada um desses gráficos representa o impacto dos custos de transação com valores de 0,10 (bps), 0,50 (bps) e 1,0 (bps), respectivamente.

Figura 5-Riqueza acumulada com custo de transação 0,10 (bps).



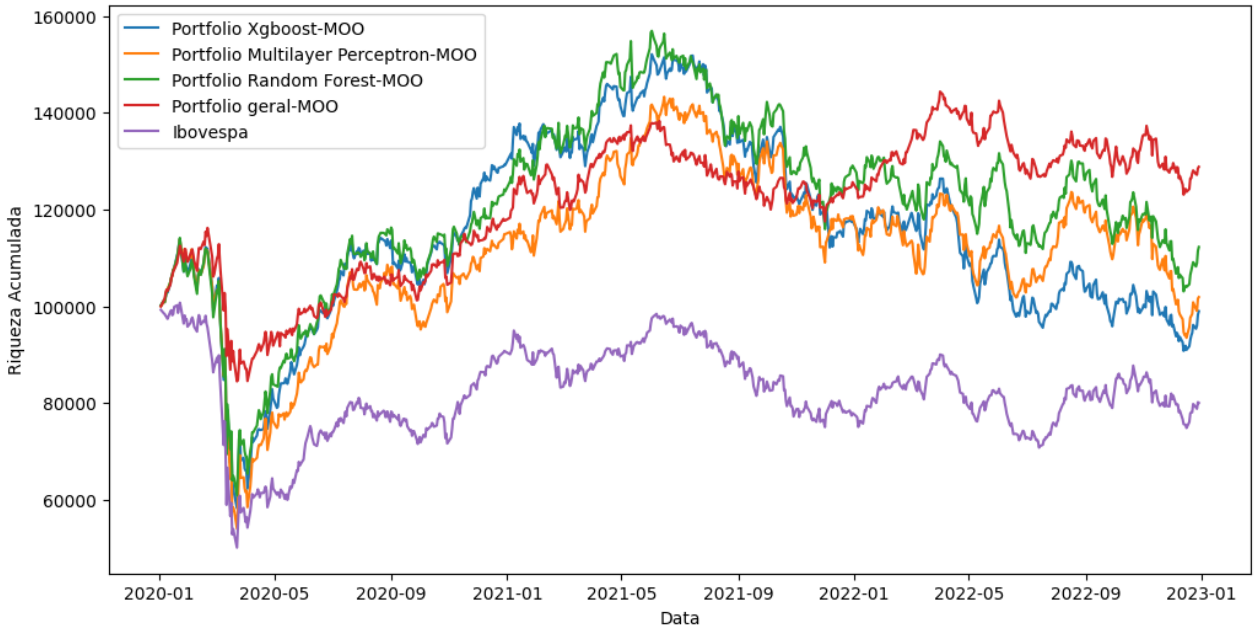

Fonte: Elaborado pelo autor.

Figura 6-Riqueza acumulada com custo de transação 0,50 (bps).




Fonte: Elaborado pelo autor.

Figura 7-Riqueza acumulada com custo de transação 1,00 (bps).




Fonte: Elaborado pelo autor.

Observa-se nas análises realizadas que, mesmo com a inclusão dos custos de transação em todos os cenários, os portfólios otimizados conseguiram superar o *benchmark* de mercado. Além disso, observa-se um padrão consistente em que o portfólio composto por todos os ativos apresenta o melhor desempenho, seguido pelos portfólios formados por *Random Forest*, *Multilayer Perceptron* e *XGBoost*. Ressalta-se que os portfólios formados pelos modelos de *machine learning* mostram-se relativamente mais eficientes, pois apresentaram um número de ativos inferior, cerca de 22 ativos, em comparação com o número de ativos da otimização do portfólio que considera todos os ativos (62 ativos), conforme apresentado na Tabela 8.

## 5 Considerações Finais

Este estudo propõe um modelo de seleção e otimização de portfólios de investimento que utiliza pré-seleção de ações por meio de modelos de *machine learning* integrada a um modelo de otimização de portfólio. O objetivo desse modelo é simultaneamente maximizar o retorno e minimizar o risco do portfólio, que é representado pela volatilidade, enquanto atende a restrições de alocação máxima e da somatória de pesos igual a 1. Além disso, foi realizada uma análise da aplicação da metodologia com e sem a presença de custos de transações.

Para a pré-seleção de ações, foram empregados os modelos de *machine learning* Xgboost, *Multilayer Perceptron* e *Random Forest*. Esses modelos foram alimentados com indicadores financeiros selecionados após revisão da literatura e passaram por um processo de *feature selection* para determinar quais são os indicadores considerados mais relevantes para a análise.

Inicialmente, os modelos de *machine learning* foram desenvolvidos e treinados utilizando a metodologia de validação cruzada. Posteriormente, esses modelos foram avaliados por meio de métricas como *recall*, precisão, acurácia e F1-*score*, com resultados similares entre todos os modelos. Com base nas previsões geradas por cada modelo, foi formado um grupo de ativos recomendados para investimento por cada um deles. Esses ativos, então, foram submetidos ao modelo de otimização multiobjetivo.

A partir da implementação do modelo de otimização, foram calculadas diferentes métricas para avaliar o desempenho dos portfólios. Em todas essas métricas, os portfólios otimizados superaram o *benchmark* de mercado. O retorno anual e o índice de Sharpe foram particularmente destacados, com os portfólios apresentando um desempenho superior. Além disso, métricas como o índice de Treynor, o beta e o Alfa de Jensen foram avaliados para os portfólios, todos indicando maior eficiência em relação ao mercado. A performance dos portfólios também foi comparada ao longo do tempo com o *benchmark* e os portfólios otimizados superaram consistentemente a carteira de mercado.

Os portfólios otimizados a partir da pré-seleção de ações com base em indicadores financeiros demonstraram resultados bastante similares. No entanto, o portfólio recomendado pelo modelo *Random Forest* se destacou ao apresentar um desempenho superior nos indicadores analisados, com um retorno médio anual de 19.11%, o que vai de encontro com a pesquisa de (MA; HAN; WANG, 2021), que apesar de ter utilizado outros indicadores financeiros e aplicar os métodos em outros mercados também obteve melhor desempenho no modelo de *machine learning Random Forest*. Considera-se que este estudo atingiu o seu objetivo e potencialmente contribui com a literatura relacionada, como trabalho de (PAIVA et al., 2019) , incluindo a *feature selection* na metodologia, uma variável *target* que é baseada no retorno do *benchmark* de mercado, e a validação cruzada, além de integrar a metodologia de *machine learning* com o modelo de otimização multibojetivo. Ademais, foram avaliados os portfólios por meio de uma variedade de indicadores de desempenho de carteiras, filtrados pelo critério de *feature selection*, o que oferece uma maior robustez à avaliação dos portfólios.

O estudo trouxe uma contribuição empírica para a literatura por meio de uma metodologia que integra um modelo de otimização multobjetivo com a capacidade preditiva dos modelos de *machine learning* de identificar investimentos por meio da pré-seleção de ações. Além disso, testa os portfólios por uma série de indicadores de desempenho, o que oferece maior robustez à metodologia do ponto de vista científico.

Apesar de o estudo atingir o objetivo proposto, ressaltam-se algumas melhorias que podem ser implementadas como sugestão de trabalhos futuros : i) testar outros modelos de *machine*

*learning* para realizar a pré-seleção de ações; ii) testar outros algoritmos para resolver o problema de otimização, como NSGA II (BARROSO; CARDOSO; MELO, 2021; DE MELO; CARDOSO; JESUS, 2022; PIMENTA et al., 2018); iii) aplicar restrição de cardinalidade aos portfólios, como em (BARROSO; CARDOSO; MELO, 2021; PAIVA et al., 2019); iv) aplicar a metodologia proposta no presente estudo a outros mercados.

## REFERÊNCIAS


ASHRAFZADEH, M. et al. Clustering-based return prediction model for stock pre-selection in portfolio optimization using PSO-CNN+MVF. **Journal of King Saud University - Computer and Information Sciences**, v. 35, n. 9, p. 101737, out. 2023.

BARROSO, B. C.; CARDOSO, R. T. N.; MELO, M. K. Performance analysis of the integration between Portfolio Optimization and Technical Analysis strategies in the Brazilian stock market. **Expert Systems with Applications**, v. 186, 2021.

BASAK, S. et al. Predicting the direction of stock market prices using tree-based classifiers. **North American Journal of Economics and Finance**, v. 47, p. 552–567, 2019.

BEHERA, J. et al. Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets. **Engineering Applications of Artificial Intelligence**, v. 120, 2023.

BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of K-fold cross-validation. **Journal of Machine Learning Research**, v. 5, p. 1089–1105, 2004.

BHUVANA CHANDRIKA, A. T.; VIJAYANAND, R.; RAO, D. P. S. ANDROID MALWARE DETECTION USING EXTRA TREES CLASSIFIER BASED FEATURE SELECTION AND MACHINE LEARNING. **international journal of techno engineering**, v. XIII, 2021.

BODNAR, T.; MAZUR, S.; OKHRIN, Y. Bayesian estimation of the global minimum variance portfolio. **European Journal of Operational Research**, v. 256, n. 1, p. 292–307, 2017.

BOONGASAME, L.; SONGRAM, P. Cryptocurrency price forecasting method using long short-term memory with time-varying parameters. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 30, n. 1, p. 435–443, 2023.

BREIMAN, L. Random Forests. **Machine Learning**, 2001.

CHANG, T.-J.; YANG, S.-C.; CHANG, K.-J. Portfolio optimization problems in different risk measures using genetic algorithm. **Expert Systems with Applications**, v. 36, n. 7, p. 10529–10537, set. 2009.

CHAWEEWANCHON, A.; CHAYSIRI, R. Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning. **International Journal of Financial Studies**, v. 10, n. 3, p. 64, 8 ago. 2022.

CHEN, S. et al. Bollinger Bands Trading Strategy Based on Wavelet Analysis. **Applied Economics and Finance**, v. 5, n. 3, p. 49, 2 abr. 2018.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: ACM, 13 ago. 2016.

CHEN, W. et al. Mean–variance portfolio optimization using machine learning-based stock price prediction. **Applied Soft Computing**, v. 100, p. 106943, mar. 2021.

CHEN, Y.; HAO, Y. Integrating principle component analysis and weighted support vector machine for stock trading signals prediction. **Neurocomputing**, v. 321, p. 381–402, dez. 2018.

DE MELO, M. K.; CARDOSO, R. T. N.; JESUS, T. A. Multiobjective Model Predictive Control for portfolio optimization with cardinality constraint. **Expert Systems with Applications**, v. 205, 2022.

DIMARZIO, F.; MATIAS FILHO, J.; FERNANDES, R. A. BEHAVIORAL FINANCE: EMPIRICAL EVIDENCE USING MAGIC FORMULA IN THE BRAZILIAN STOCK MARKET. **RAM. Revista de Administração Mackenzie**, v. 21, n. 6, 2020.

FAVERO; BELIFIORE. **Manual de análise de dados** . Grupo GEN LTC ed. [s.l: s.n.].

FERRI, F. J. et al. **Comparative Study of Techniques for Large-Scale Feature Selection**. [s.l: s.n.]. Disponível em: <The Pennsylvania State University>. Acesso em: 3 maio. 2023.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2 abr. 2006.

GUERARD, J. B.; MARKOWITZ, H.; XU, G. Earnings forecasting in a global stock selection model and efficient portfolio construction and management. **International Journal of Forecasting**, v. 31, n. 2, p. 550–560, abr. 2015.

GUIMARÃES JÚNIOR, F. R. F.; CARMONA, C. U. DE M.; GUIMARÃES, L. G. DE A. CARTEIRAS FORMADAS POR MEIO DE VARIÁVEIS FUNDAMENTALISTAS APRESENTAM BOM DESEMPENHO DE MERCADO? **Gestão & Regionalidade**, v. 31, n. 91, 2 abr. 2015.

HUANG, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. **Applied Soft Computing**, v. 12, n. 2, p. 807–818, fev. 2012a.

HÜBNER, G. The Generalized Treynor Ratio. **Review of Finance**, v. 9, n. 3, p. 415–435, 1 jan. 2005.

JENSEN, M. C. THE PERFORMANCE OF MUTUAL FUNDS IN THE PERIOD 1945-1964. **The Journal of Finance**, v. 23, n. 2, p. 389–416, maio 1968.

KONG, H.; YUN, W.; KIM, W. C. Tracking customer risk aversion. **Finance Research Letters**, 2023.

KRAFT, D. A Software Package for Sequential Quadratic Programming. **DLR German Aerospace Center – Institute for Flight Mechanics**, p. 88–28, 1998.

LI, L. **Selecting Portfolios Directly Using Recurrent Reinforcement Learning**. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. **Anais**...2020.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, 2021.

MANJUNATH, C.; MARIMUTHU, B.; GHOSH, B. Analysis of Nifty 50 index stock market trends using hybrid machine learning model in quantum finance. **International Journal of Electrical and Computer Engineering**, v. 13, n. 3, p. 3549–3560, 2023.

MAZRAEH, N. B. et al. Stock Portfolio Optimization Using a Combined Approach of Multi Objective Grey Wolf Optimizer and Machine Learning Preselection Methods. **Computational Intelligence and Neuroscience**, v. 2022, p. 1–20, 29 ago. 2022.

NOBRE, J.; NEVES, R. F. Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. **Expert Systems with Applications**, v. 125, p. 181–194, jul. 2019.

ORIMOLOYE, L. O. et al. Comparing the effectiveness of deep feedforward neural networks and shallow architectures for predicting stock price indices. **Expert Systems with Applications**, v. 139, p. 112828, jan. 2020.

ORRA, A.; SAHOO, K.; CHOUDHARY, H. **Machine Learning-Based Hybrid Models for Trend Forecasting in Financial Instruments**. [s.l: s.n.]. v. 547

PADHI, D. K. et al. An Intelligent Fusion Model with Portfolio Selection and Machine Learning for Stock Market Prediction. **Computational Intelligence and Neuroscience**, v. 2022, 2022.

PAIVA, F. D. et al. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. **Expert Systems with Applications**, v. 115, p. 635–655, 2019.

PIMENTA, A. et al. An Automated Investing Method for Stock Market Based on Multiobjective Genetic Programming. **Computational Economics**, v. 52, n. 1, p. 125–144, 2 jun. 2018a.

PRASTYO, P. H.; ARDIYANTO, I.; HIDAYAT, R. **A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods**. Proceedings - 2020 6th International Conference on Science and Technology, ICST 2020. **Anais**...2020.

RUBESAM, A.; BELTRAME, A. L. Carteiras de Variância Mínima no Brasil. **Brazilian Review of Finance**, v. 11, n. 1, 2013.

SAKHARE, N. N.; SHAIK, I. S.; SAHA, S. Prediction of stock market movement via technical analysis of stock data stored on blockchain using novel History Bits based machine learning algorithm. **IET Software**, 12 jan. 2023.

SHARMA, J. et al. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. **EURASIP Journal on Information Security**, v. 2019, n. 1, p. 15, 22 dez. 2019.

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

SON, D.; PARK, J.; HUH, E.-N. **Dynamic SAR for Efficient Container Auto-Scaling Based on Network Traffic Prediction**. TIMES-iCON 2018 - 3rd Technology Innovation Management and Engineering Science International Conference. **Anais**...2019.

SONG, X.-P. et al. Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China. **Journal of Forecasting**, v. 33, n. 8, p. 611–626, dez. 2014.

SILVA, N. F. et al. An integrated CRITIC and Grey Relational Analysis approach for investment portfolio selection. **Decision Analytics Journal**, p. 100285, jul. 2023.

SILVA, N. F. et al. Portfolio optimization based on the pre-selection of stocks by the Support Vector Machine model. **Finance Research Letters**, v. 61, p. 105014, mar. 2024.

TENG, T.; MA, L. Deep learning-based risk management of financial market in smart grid. **Computers and Electrical Engineering**, v. 99, p. 107844, abr. 2022.

THENMOZHI, M.; SARATH CHAND, G. Forecasting stock returns based on information transmission across global markets using support vector machines. **Neural Computing and Applications**, v. 27, n. 4, p. 805–824, 11 maio 2016.

TIN KAM HO. **Random decision forests**. Proceedings of 3rd International Conference on Document Analysis and Recognition. **Anais**...IEEE Comput. Soc. Press, 1995.

VALLE, C. A.; MEADE, N.; BEASLEY, J. E. Absolute return portfolios. **Omega**, v. 45, p. 20–41, jun. 2014.

WANG, W. et al. Portfolio formation with preselection using deep learning from long-term financial data. **Expert Systems with Applications**, v. 143, p. 113042, abr. 2020.

WIDIASARI, I. R.; NUGROHO, L. E.; WIDYAWAN. **Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining**. 2017 International Conference on Innovative and Creative Information Technology (ICITech). **Anais**...IEEE, nov. 2017.

WU, W. et al. Dynamic mean-downside risk portfolio selection with a stochastic interest rate in continuous-time. **Journal of Computational and Applied Mathematics**, v. 427, p. 115103, ago. 2023.

ZHOU, F. et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. **Applied Soft Computing**, v. 84, p. 105747, nov. 2019.

## PRODUTO 4 (Técnico/tecnológico): Algoritmo de seleção e otimização de portfólio de investimento

### 1. Introdução

O processo de seleção e otimização de investimentos é uma tarefa complexa que envolve uma variedade de alternativas e critérios (SILVA et al., 2023). Para determinar em quais ativos investir e qual proporção alocar em cada um, pesquisadores têm aplicado diversas metodologias com o objetivo de aprimorar os resultados. Isso inclui o uso de técnicas de *machine learning* para a pré-seleção de ativos (CHAWEEWANCHON; CHAYSIRI, 2022; MA; HAN; WANG, 2021; PAIVA et al., 2019) e abordagens de otimização multobjetivo (ASGARI; BEHNAMIAN, 2022; ZHANG; LU, 2021). Essas estratégias visam a maximização dos retornos e a minimização dos riscos, proporcionando uma abordagem mais eficiente e fundamentada na tomada de decisões de investimento.

No entanto, além de uma abordagem robusta que considere os múltiplos critérios relacionados ao mercado na seleção de investimentos, torna-se necessario o uso de uma ferramenta computacional adequada para auxiliar os investidores na tomada de decisões. A implementação de um modelo que utiliza *machine learning* para analisar quais ativos investir, por meio de indicadores financeiros, e otimização multobjetivo para alocar os investimentos pode representar um desafio significativo do ponto de vista do investidor. Dada a complexidade desses métodos, é crucial contar com uma ferramenta intuitiva e acessível que traduza esses modelos complexos em informações compreensíveis para os investidores.

Dessa forma, este produto tecnológico procurou compilar a metodologia apresentada no produto bibliográfico 3, por meio de um algoritmo desenvolvido na linguagem Python. Esse algoritmo possibilitará que o investidor aplique a metodologia proposta ou que a pesquisa seja replicada em outros mercados. A seguir será apresentado a metodologia do algoritimo.

### 2. Metodologia

O algoritmo desenvolvido podera utilizado em ambiente local, por meio de um gerador de código, quanto no ambiente de colaboração do Google Colab. Este algoritmo está dividido em três partes distintas: i) Coleta, tratamento, Cálculo de indicadores financeiros e seleção de alternativas de investimentos com machine learning; ii) implementação do modelo multiobjetivo e análise de desempenho das carteiras; e implementação do modelo multiobjetivo com análise da inserção de custos de transação;

Ao oferecer essa estrutura modular, o algoritmo proporciona flexibilidade aos usuários, permitindo a escolha de partes específicas conforme suas necessidades ou a execução completa como ocorreu na metodologia do produto bibliográfico 3. Essa versatilidade visa atender a diferentes contextos e requisitos na seleção e otimização de investimentos.

A seguir listamos as principais etapas da primeira parte do algoritmo:

1. Coleta de Dados:

- Definição dos ativos para análise.

- Download dos dados históricos, incluindo preço de abertura, preço máximo, preço mínimo, preço de fechamento, preço ajustado, volume e retorno logarítmico.

2. Cálculo de indicadores financeiros:

- Implementação de indicadores como Média Móvel Simples (SMA), Média Móvel Exponencial (EMA), Bandas de Bollinger (BB), Parabolic SAR (PSAR), On-Balance Volume (OBV), MACD, RSI, Estocástico e Momentum.

3.Definição do Benchmark (Ibovespa):

- Download dos dados do benchmark.

4. Preparação dos Dados para Modelagem:

- Manipulação dos dados para facilitar a análise, incluindo o cálculo da diferença percentual entre os retornos dos ativos e do benchmark.

5. *Feature Engineering*:

- Criação de indicadores adicionais a partir dos dados, como média móvel simples (SMA) e outras características relevantes.

6. Limpeza e Filtragem de Dados:

- Exclusão de ativos com dados ausentes e normalização do conjunto de dados.

7. *Feature Selection*:

- Utilização de métodos como *Extra Trees Classifier* para selecionar as características mais importantes para a modelagem.

8. Treinamento de Modelos:

- Implementação e treinamento de modelos de machine learning, incluindo *XGBoost, Multilayer Perceptron e Random Forest*.

9. Avaliação dos Modelos:

- Utilização de métricas como precisão, acurácia, F1-Score e recall para avaliar o desempenho dos modelos.

10. Seleção de Ativos para Investimento:

- Aplicação de limiares personalizados para filtrar ativos recomendados pelos modelos.

11.Avaliação de Consistência nas Recomendações:

- Análise da frequência de aparições dos ativos recomendados para garantir consistência nas escolhas.

12. Apresentação dos Resultados:

- Criação de tabelas resumindo os ativos recomendados por cada modelo e métricas de desempenho.

A partir da primeira parte do algoritmo, observa-se que o processo abrange todas as etapas desde a coleta de dados até o cálculo de indicadores financeiros e a seleção de investimentos por meio de *machine learning*. Agora, na segunda parte do algoritmo, destacaremos as etapas relacionadas à otimização multiobjetivo e à análise de desempenho das carteiras.

1. Coleta de Dados:

- Inicialmente, são definidos os ativos que serão analisados no processo.

- Os dados históricos de preços ajustados desses ativos são baixados por meio do download de informações financeiras.

2 . Cálculo dos Retornos Logarítmicos para cada modelo:

- São calculados os retornos logarítmicos diários para cada ativo, refletindo as variações percentuais nos preços.

3. Implementação do modelo de otimização multiobjetivo:

- O modelo de otimização é projetado para buscar uma alocação de ativos que maximize o retorno e minimize o risco simultaneamente.

- Restrições são aplicadas, limitando a alocação máxima permitida para cada ativo.

4. Cálculo do modelo de otimização:

- A otimização do portfólio é realizada utilizando o modelo previamente implementado.

5. Análise de Desempenho das Carteiras:

- As carteiras otimizadas são avaliadas com base em diversas métricas, incluindo o índice de Sharpe, índice de Treynor, coeficiente beta, Alfa de Jensen, prêmio pelo risco e VAR (Value at Risk).

- A performance das carteiras é analisada graficamente, permitindo uma comparação visual ao longo do tempo.

- As carteiras otimizadas também são comparadas com carteiras igualmente ponderadas, proporcionando uma perspectiva adicional.

Essas etapas compõem o fluxo geral do processo do algoritmo da parte 2, desde a coleta de dados até a análise de desempenho das carteiras. A parte 3 do algoritmo assemelha-se à parte 2, com a distinção crucial da inclusão dos custos de transação no processo de otimização. Nesta etapa, ao

ocorrer uma alteração nos pesos do modelo, é calculado um custo de transação que impacta os resultados dos modelos, proporcionando uma abordagem mais realista e financeiramente sensível.

## 3. Considerações Finais

O algoritmo desenvolvido neste produto técnico apresenta uma estrutura modular que o torna versátil para diferentes usuários, sendo aplicável tanto por pesquisadores quanto por investidores. A modularidade permite que partes específicas, como o módulo de otimização ou a seleção de investimentos com *machine learning*, sejam utilizadas de forma independente, atendendo a diversas necessidades e cenários de aplicação.

Como recomendação para pesquisas futuras, sugere-se o desenvolvimento de uma interface para o algoritmo, tornando-o acessível de forma online. Essa iniciativa visa aprimorar a usabilidade para investidores, proporcionando uma plataforma web intuitiva e de fácil navegação.

## REFERENCIAS


ASGARI, H.; BEHNAMIAN, J. Multi-objective stock market portfolio selection using multi-stage stochastic programming with a harmony search algorithm. **Neural Computing and Applications**, v. 34, n. 24, p. 22257–22274, 2022.

CHAWEEWANCHON, A.; CHAYSIRI, R. Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning. **International Journal of Financial Studies**, v. 10, n. 3, p. 64, 8 ago. 2022.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, 2021.

PAIVA, F. D. et al. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. **Expert Systems with Applications**, v. 115, p. 635–655, 2019.

SILVA, N. F. et al. An integrated CRITIC and Grey Relational Analysis approach for investment portfolio selection. **Decision Analytics Journal**, p. 100285, jul. 2023.

ZHANG, Y.; LU, S. **Multi-model fusion method and its application in prediction of stock index movements**. ACM International Conference Proceeding Series. **Anais**...2021.