Uma Estratégia para Alocação de Carteira de Ações usando Algoritmos de Aprendizado de Máquina e Regras Fuzzy

Mario Gambim [1], **Heloisa A. Camargo** [1], **Giancarlo Lucca** [2]
Graçaliz Dimuro [3], **Tiago Asmus** [4]

[1]Departamento de Computação – UFSCar – São Carlos – SP – Brasil

[2]Centro de Ciências Sociais e Tecnológicas – UCPEL – Pelotas – RS – Brazil

[3]Centro de Ciências Computacionais – FURG – Rio Grande – RS – Brasil

[4]Instituto de Matemática, Estatística e Física – FURG – Rio Grande – RS – Brasil

mariolg@estudante.ufscar.br, heloisacamargo@ufscar.br,

giancarlo.lucca88@gmail.com, gracalizdimuro@furg.br,

tiagoasmus@furg.br

Abstract. *In the stock market, investors interested in long-term investments look for ways to select stocks that have the potential for future appreciation to improve their earnings. Machine learning methods have a long history of applications in the financial market. This article presents and evaluates a simple and effective strategy for allocating stock portfolios, which uses machine learning algorithms to predict the performance of stocks and select a few of them to compose the portfolio. The proposed method is particularly useful for investors with little knowledge of the market. The results obtained show that the proposed strategy led to the allocation of portfolios with higher profits than those of the S&P500 market index and those of a manual method known in the literature.*

Resumo. *No mercado de ações, investidores interessados em investimentos de longo prazo buscam formas de selecionar ações com potencial de valorização futura para melhorar seus rendimentos. Os métodos de aprendizado de máquina têm um longo histórico de aplicações no mercado financeiro. Este artigo apresenta e avalia uma estratégia simples e eficaz para alocação de carteiras de ações, que utiliza algoritmos de aprendizado de máquina para prever o desempenho das ações e selecionar algumas delas para compor a carteira. O método proposto é particularmente útil para investidores com pouco conhecimento do mercado. Os resultados obtidos mostram que a estratégia proposta levou à alocação de carteiras com lucros superiores aos do índice de mercado S&P500 e aos de um método manual conhecido na literatura.*

1. Introdução

A constante evolução do mercado financeiro ao longo dos anos tornou esse segmento mais robusto e complexo no que diz respeito às operações envolvidas, como também gerou uma maior gama de aplicações, que podem ser de renda fixa ou variável. Com essa abundância de opções de aplicações os investidores têm o desafio de procurar por aquelas

que tenham um maior potencial de valorização de acordo com sua estratégia de investimento, que pode ser de curto, médio ou longo prazo. Para isso, buscam informações ou métodos que sejam capazes de apontar as melhores aplicações. Diante disso, métodos computacionais passaram a ser aplicados por investidores com o objetivo de encontrar as melhores opções. Dentre eles podemos destacar métodos de Aprendizado de Máquina (AM), como no trabalho de [Castro 2009] que utilizou a lógica fuzzy para prever, dentre um conjunto de ações, aquelas que teriam maior potencial de valorização futura, o trabalho de [Franchi 2021] que fez uso de aprendizado por reforço para otimização de carteiras de investimento e [Santos Junior 2015] que utilizou máquinas de vetores de suporte para predição dos valores dos preços de ações.

Para aplicações de longo prazo, o investidor deseja selecionar ações de empresas para construir uma carteira de investimentos, buscando informações de diversas fontes e recorrendo ao uso de ferramentas automatizadas. Ao contrário de aplicações a curto e médio prazo, que são tratadas com o uso de séries temporais, as aplicações de longo prazo requerem métodos mais simples, já que os dados são tratados em lote e não há necessidade de tratar questões relacionadas à dinâmica temporal, como sazonalidades e ruídos nos dados.

Estudos como o de [Basu 1983] identificaram múltiplos de mercado capazes de explicar o retorno futuro das ações, contrariando a teoria dos mercados eficientes, a qual diz que qualquer informação disponível é instantaneamente assimilada pelos operadores de mercado, impedindo que qualquer investidor consiga vantagem na obtenção de lucros [Ross 2009]. Uma análise fundamentalista apresentada por [Graham 2016] determina que os múltiplos P/L (Preço/Lucro) e P/VP (Preço/Valor Patrimonial) fornecem informações suficientes para que o investidor tome suas decisões para aplicações de longo prazo. Tais múltiplos podem indicar se o preço das ações de um empresa estão descontados, uma vez que o P/L mostra a razão entre o valor das ações da empresa e seu lucro total e o P /VP é a razão entre o preço das ações e o valor patrimonial da empresa. Assim, ações com baixo P/L indicam que uma empresa tem um alto lucro frente ao valor de suas ações e ações com baixo P/VP indicam que uma empresa tem alto lucro frente ao seu valor patrimonial.

Neste trabalho foi elaborada e avaliada uma estratégia para alocação de carteiras para aplicações de longo prazo (1 ano) com o uso de AM, de acordo com a teoria de análise fundamentalista, que visa oferecer um meio simples e eficaz para o investidor decidir sobre as ações que serão incluídas na carteira. Nessa linha, algoritmos de regressão foram aplicados aos dados para predizer o desempenho futuro das ações no período de um ano e, com base nessa predição, selecionar as ações mais promissoras. Como atributos de entrada foram utilizados os dois múltiplos P/L e P/VP e o volume de negociação, calculado como o número de negociações de uma determinada ação no período de um ano, assim podendo indicar a liquidez daquele ativo. Já o atributo de saída desejado é o desempenho, ou seja, a predição da variação do valor do preço da ação no período de um ano. Para isso modelos de regressão foram treinados para predizer o desempenho de ações no fechamento de um ano a partir dos valores dos três atributos na abertura do ano. Os modelos escolhidas foram: Regressão Linear (RL), Árvore de Regressão (AR), Random Forest (RF), K-Nearest Neighbors (KNN), Baeysian Ridge (BR) e Sistemas de Inferência Fuzzy (SIF). Foram avaliados os desempenhos dos modelos comparando resultados das predições com dados de treinamento e teste. Depois foram avaliadas e as carteiras obtidas

com cada método de predição. As análises mostram que o desempenho das carteiras geradas pela estratégia proposta apresenta lucros superiores aos do índice de mercado S&P500 e aos de um método manual conhecido na literatura.

2. Conceitos Básicos e Trabalhos Relacionados

Nesta seção serão apresentados conceitos básicos necessários para compreensão do trabalho, tanto da área de alocação de carteiras de ações quanto da área de AM. Também serão apresentados trabalhos que abordam problemas semelhantes ao descrito aqui.

2.1. Alocação de Carteiras de Ações

No mercado financeiro, os investidores buscam, dentre uma gama de investimentos, aqueles que tem a chance de trazer maior lucro no futuro. No caso do mercado de ações os investidores selecionam os ativos que consideram mais promissores e montam uma carteira de ações. As diversas teorias de mercado desenvolvidas até hoje tentam explicar o funcionamento dos mercados, ou seja, como os agentes econômicos se comportam frente as situações impostas pelo cotidiano. Investidores procuram formular suas estratégias de investimentos baseados nessas teorias, para que assim possam maximizar seus ganhos e minimizar seus prejuízos. Duas dessas teorias baseiam-se na Hipótese dos Mercados Eficientes e na Análise Fundamentalista.

A Hipótese dos Mercados Eficientes de [Fama 1970] diz que os preços dos ativos financeiros refletem todas as informações do mercado e que o preço de um ativo é o real valor daquele ativo diante de um conjunto de informações já disponíveis, ou seja assume que o preço de um ativo no momento da compra e seu valor, definido pelas características da corporação representada por esse ativo, como governança, endividamento, setor de atuação, dentre outras, são iguais. No entanto, Hipótese dos Mercados Eficientes também considera que eventualmente preço e valor possam assumir valores diferentes, mas essas diferenças seriam aleatórias e passageiras, não sendo possível tirar vantagem dessas situações. Outro ponto levantado por essa hipótese é de que todos os agentes econômicos tomam decisões racionais e são avessos a riscos excessivos. Sendo assim, os investidores não devem levar em conta situações de cunho emocionais e efêmeras, escolhendo ativos baseando-se no fato de que o mercado precifica um ativo com seu valor justo.

Análise fundamentalista é um método de avaliação sobre a situação financeira, econômica e setorial de uma empresa e para isso são usados múltiplos que podem trazer uma perspectiva sobre a situação de determinada corporação. Contrariando a Hipótese dos Mercados Eficientes, [Basu 1977] e [Graham 2016] mostraram que há dois múltiplos fundamentalistas de mercado capazes de indicar se uma ação está com seu preço subvalorizado ou supervalorizado, permitindo assim que o investidor possa encontrar essas assimetrias e adquirir ações com um grande potencial de valorização, como também evitar a compra de ações que tenham seu preço supervalorizado.

Dentre esses múltiplos fundamentalistas, um dos mais utilizados é o P/L, que é a razão entre o preço de mercado de uma ação e o lucro líquido por ação mais atual referente ao período de um ano, dividido pelo número de ações existentes para aquele empresa. O estudo de [Basu 1977] analisou ações negociadas na Bolsa de Valores de Nova York (NYSE) e mostrou que ações com baixo P/L tinham maior potencial de valorização do que as de alto P/L. Outros estudos mostraram que a análise de uma ação baseada somente no

P/L não é eficiente, uma vez que outras características como valor de mercado da empresa e setor de atuação têm ligação direta com o P/L. De acordo com [Graham 2016], ações negociadas a preços menores do que o valor patrimonial da empresa são subvalorizadas e as que apresentam preços maiores são supervalorizadas. Nesse sentido, outro indicador fundamentalista amplamente utilizado é P/VP, que é a razão entre o preço de uma ação e o valor patrimonial da empresa, o que sugere que o investidor deve considerar empresas com P/VP baixos no momento de montar sua carteira de ações.

2.2. Regressão

A regressão é um tipo de AM supervisionado que, quando recebe um conjunto de dados de treinamento, gera um modelo que pode ser interpretado como a aproximação de uma função que prevê valores contínuos. Existe uma grande quantidade de algoritmos de regressão, que se adequam a diferentes problemas, levando em conta os tipos, organização e estrutura do conjunto de dados. Na sequência vamos descrever brevemente os métodos selecionados para este trabalho.

- **Regressão Linear (RL)** - Estão entre os mais difundidos modelos de regressão. Nesses modelos, dado um conjunto de dados espera-se que o valor destino seja a combinação linear dos atributos de entrada. Quando temos vários atributos no conjunto de dados, chamamos de regressão linear múltipla.
- **Árvores de decisão e regressão** - Estão entre os métodos mais utilizados para AM supervisionado. Os modelos chamados árvores de decisão são designados para problemas de classificação, enquanto os chamados árvore de regressão (AR) para problemas de regressão, no entanto a interpretação dos modelos e algoritmos para ambas árvores são muito semelhantes e sendo assim usaremos o termo árvore de decisão como uma forma genérica para ambos os modelos [Faceli et al. 2011]. A construção de uma árvore de decisão começa pela raiz e, a cada nó criado, um atributo é selecionado para esse nó, o que pode ser feito por diferentes índices, como ganho de informação, gini, dentre outros. Cada nó folha terá uma classe para problemas de classificação e um valor numérico ou equação para problemas de regressão.
- **Random Forest (RF)** - É uma generalização das árvores de decisão tradicionais, podendo ser utilizado para classificação ou regressão. Sua implementação tem como premissa a combinação de um conjunto de árvores de decisão com a flexibilidade e aleatoriedade para melhor precisão, em que cada árvore será utilizada na escolha do resultado final. O RF é eficiente para lidar com um conjunto de dados grande por ser mais rápido e ter como principal objetivo minimizar o sobreajuste, que é quando uma função se ajusta muito bem ao conjunto de dados de treinamento e mostra-se ineficaz para prever novos resultados. [Han et al. 2012]. O processo do RF consiste na seleção aleatória de um subconjunto dos dados originais, seguida da seleção também de maneira aleatória das caraterísticas para montagem das árvores, sendo que diversas árvores serão criadas a partir de subconjuntos diferentes. Para predizer a saída, dada uma nova instância de dado, cada árvore é percorrida para definir sua saída e o valor final é definido com base na saída de todas as árvores.
- **K-Nearest Neighbors (KNN)** - É um método de AM da categoria de aprendizado *lazy learning*, no qual não é feita a generalização do conhecimento e o modelo é

formado pelas próprias instâncias, que são armazenadas. Os valores dos atributos de entrada são representados por tuplas com n atributos. Quando recebe-se uma nova tupla desconhecida, o classificador KNN procura no espaço de padrões pelas k tuplas de treinamento que estão mais próximas a tupla desconhecida, que são os *K-Vizinhos-Mais-Próximos*. A proximidade entre as tuplas de treinamento e a desconhecida é definida por uma métrica de distância, podendo esta ser a euclidiana. O valor predito será calculado pela média ou mediana dos valores de saída das instâncias mais semelhantes. [Han et al. 2012].

- **Bayesian Ridge (BR)** - O regressor Ridge é uma versão regularizada da RL, em que uma regularização é adicionada a função de custo. Isso faz com que o algoritmo de aprendizado não ajuste os dados e mantenha os pesos do modelo tão pequenos quanto possível. É importante salientar que a regularização é adicionada apenas durante o treinamento, uma vez que com o modelo treinado deseja-se usar uma medida de desempenho não regularizada para avaliar o desempenho do modelo [Géron 2019]. Uma interpretação possível do ridge regressor adota o ponto de vista Bayesiano, caso em que o algoritmo é denominado Bayesian Ridge (BR) [Tipping 2001] [MacKay 1992].

2.3. Sistemas de Inferência Fuzzy

Os conjuntos fuzzy foram apresentados inicialmente por [Zadeh 1965] com o objetivo de replicar a lógica do pensamento humano em lidar com processos complexos, que são baseados informações imprecisas e fronteiras mal definidas. Os conjuntos fuzzy buscam tratar situações que fogem a teoria clássica dos conjuntos, em que um elemento de um universo pertence ao conjunto ou não pertence ao conjunto. No cotidiano, os conjuntos clássicos não são capazes de lidar com determinados tipos de categorias, pois não apresentam flexibilidade para lidar com situações de pertinência que não sejam a absoluta. Diante disso, os conjuntos fuzzy permitem a representação de categorias com limites imprecisos e que seja possível trabalhar com elementos que estejam na fronteira, e esses possam assumir um pertinência parcial a um ou mais conjuntos [Klir and Yuan 1995].

Os conjuntos fuzzy são definidos por uma função que generaliza a função característica dos conjuntos clássicos, chamada de função de pertinência. Pode-se denotar a função de pertinência como $A : X \rightarrow [0, 1]$ em que $A(x)$ é o grau de pertinência de $x \in X$ em A. A função de pertinência conecta os elementos do conjunto universo X a valores reais do intervalo $[0, 1]$,sendo 0 para nenhum grau de pertinência e 1 para pertinência total. As funções de pertinência podem assumir diversas formas, sendo as mais comuns a triangular, trapezoidal e gaussiana [Klir and Yuan 1995]. Nesse trabalho, faremos uso da função triangular.

O Sistema de Inferência Fuzzy (SIF) é baseado nos conceitos de conjuntos, regras e raciocínio fuzzy. As regras se mostram eficientes na modelagem de proposições em linguagem natural e esses sistemas podem ser encontrados de forma isolada ou combinada com outros métodos [Das et al. 2022] em aplicações em áreas de classificação, regressão e agrupamento de dados [Skrjanc et al. 2019], predição de séries temporais [Orang et al. 2022], tomada de decisões [Bisht and Kumar 2022] entre outros. O SIF é um modelo que faz predições numéricas a partir dos dados de entrada e, por esse motivo é utilizado neste trabalho como um método de regressão. A principal motivação para incluir o SIF entre os algoritmos selecionados é a perspectiva de investigar, em conti-

nuidade a esta pesquisa, como a interpretabilidade das regras fuzzy pode ser explorada para corroborar a teoria fundamentalista adotada na estratégia abordada neste trabalho [Mendel and Bonissone 2021].

O SIF é composto pelos seguintes componentes: base de regras, que contém um conjunto de regras fuzzy; base de dados, que define as funções de pertinência utilizadas nas regras fuzzy; e o mecanismo de inferência, que executa a inferência e retorna uma saída. Nesse trabalho utilizaremos o modelo de Mamdani [Mamdani and Assilian 1975], que é um método de inferência composicional simplificada que utiliza conjuntos fuzzy nos antecedentes e nos consequentes das regras fuzzy. Esse modelo recebe entradas numéricas, que são os valores dos atributos de entrada da instância para a qual se quer fazer uma predição. As pertinências desses valores aos respectivos conjuntos dos antecedentes das regras são calculadas e combinadas pelo operador mínimo, obtendo-se assim o grau de disparo de cada regra. As regras com grau de disparo maior que zero vão gerar uma saída que é um conjunto fuzzy obtido a partir do conjunto fuzzy do consequente da regra, calculando-se o menor valor entre o grau de pertinência de cada elemento do domínio da variável de saída nesse conjunto e o grau de disparo da regra. A saída é um conjunto fuzzy que resulta da agregação do conjunto fuzzy produzido pela inferência de cada regra. Para obter-se uma saída final não fuzzy é aplicado um método de defuzificação. O SIF é um modelo que faz predições numéricas a partir dos dados de entrada e, por esse motivo é utilizado neste trabalho como um método de regressão.

2.4. Trabalhos relacionados

Uma estrutura integrada de tomada de decisão multicritério usando conjunto fuzzy triangulares é desenvolvida para a construção de portfólio, unificando as avaliações de um investidor iniciante e de um especialista no mercado de ações em [Bisht and Kumar 2022]. O trabalho de [Jiménez-Preciado et al. 2022] teve como objetivo desenvolver um modelo de AM para detectar empresas com vantagens competitivas duradouras de acordo com seus índices financeiros, a fim de melhorar o desempenho das carteiras de investimento. Após calcular os índices financeiros das empresas pertencentes ao S&P500, uma avaliação quanto às vantagens competitivas da empresa é atribuída a cada índice (definida entre 0 e 5), finalizando com a classificação das empresas em três categorias. Finalmente, vários modelos de ML para classificação são aplicados, visando obter um método eficiente, mais rápido e menos dispendioso para selecionar empresas com vantagens competitivas duradouras. Ambos os trabalhos citados têm, como o apresentado neste artigo, o objetivo de simplificar o processo de seleção de ações para investidores inexperientes. Entretanto, esses dois trabalhos requerem o uso de procedimentos mais complexos, seja para obter e estruturar o conhecimento de especialistas, como em [Bisht and Kumar 2022], ou para obter dados, calcular índices e fazer a classificação manual das ações em categorias, como em [Jiménez-Preciado et al. 2022]. Já a abordagem proposta por nós requer apenas dados disponibilizados publicamente que podem ser obtidos com razoável facilidade.

Em [Hao et al. 2023] os autores tratam a tarefa de gerenciamento de portifólios por uma abordagem distinta da utilizada neste artigo, considerando algoritmos de Aprendizado por Reforço. O objetivo é treinar um agente com informações de mercado para que ele aprenda estratégias de negociação e vença o índice de mercado sem precisar fazer nenhuma previsão sobre os movimentos do mercado. A abordagem não pressupõe nenhum conhecimento de negociação, portanto, o agente aprenderá apenas conduzindo

negociações com dados históricos.

Por fim, os trabalhos de [Santos Junior 2015] e [Mazraeh et al. 2022] apresentam propostas de uso de AM para gerar ou otimizar portifólios de ações, embora adotando abordagens diferente da explorada no trabalho desenvolvido por nós, uma vez que trata o problema de previsão usando dados de séries temporais.

3. Estratégia para Alocação de Carteira de Ações

Nesta seção será descrita a estratégia proposta para alocação de carteiras de ações utilizando algoritmos e regressão e seguindo a análise fundamentalista que defende que os múltiplos P/L e P/VP oferecem informações suficientes para que o investidor possa tomar decisões sobre seus investimentos a longo prazo de maneira simples e eficaz.

3.1. Coleta e Análise Exploratória de Dados

Foram utilizadas ações que compõem o índice S&P500, pertencente ao mercado americano, no ano de 2022. O índice S&P500 é um índice composto por 500 ações cotadas nas bolsas de valores de Nova York NYSE (The New York Stock Exchange) e NASDAQ (National Association of Securities Dealers Automated Quotations), qualificados devidos ao seu tamanho de mercado, sua liquidez e sua representação de grupo industrial [Indices 2016].

Para a construção dos modelos de previsão, são considerados os atributos P/L, P/VP e VOL como variáveis de entrada e o atributo Desempenho como variável de saída:

- **P/L**: $\frac{Preco}{Lucro}$;
- **P/VP**: $\frac{Preco}{ValorPatrimonial}$;
- **VOL**: quantidade de negociações de cada ação no período de um ano.
- **Desempenho**: diferença entre os valores de fechamento e abertura de cada ação no período de um ano;

Os indicadores P/L e P/VP foram obtidos na base de dados da Bloomberg (www.bloomberg.com) enquanto que os valores de Desempenho e VOL foram retirados do site Barchart (www.barchart.com).

Na análise exploratória foram consideradas a dispersão e correlação entre os atributos. Pela dispersão concluiu-se que os dados possuem distribuição semelhante ao longo dos anos. Foi possível observar também que todos os pares de atributos têm correlação próxima de zero, o que indica que não há forte dependência entre eles. Como decorrência dessas análises, foi decidido manter os quatro atributos apresentados acima. Detalhes adicionais sobre essa análise podem ser encontrados em [Gambim 2022].

3.2. Tratamento dos Dados

Os dados coletados são referentes aos anos de 2012 a 2021, com os quais foram montadas 9 bases de dados para os anos de 2013 a 2021, já que, para o atributo de entrada VOL foi utilizado o volume de negociações referentes ao ano anterior. Para os atributos de entrada P/L e P/VP foram utilizados os valores referentes a janeiro de cada ano. Para o atributo de saída, Desempenho, foi calculado o percentual da diferença entre os valores de abertura (janeiro) e fechamento (dezembro) do mesmo ano.

O índice S&P500 é dinâmico, ou seja, novas ações podem entrar ou sair do portfólio, caso uma empresa passe a cumprir ou deixe de cumprir os requisitos para compor o índice. Sendo assim, iremos analisar empresas que se mantiveram no índice durante todo o período considerado, de 2012 a 2021. Após retirar também as ações com valores nulos em algum dos atributos, obtivemos um total de 302 ações para cada base de dados. Para fins de simplificação, chamamos esse conjunto de 302 ações de 'S&P500 Ajustado'.

As bases de dados foram organizadas de forma a gerar dois cenários para os experimentos e analisar as saídas por duas visões diferentes. No primeiro cenário, foram usadas as bases criadas como descrito anteriormente, com dados de um ano para treinamento e dados do ano seguinte para teste. Para o segundo cenário foram concatenadas bases de três anos, usadas para treinamento, e dados do ano seguinte para teste. Por exemplo, concatenou-se os dados referentes aos anos de 2013, 2014 e 2015, totalizando 906 dados para treinamento e usou-se os dados do ano 2016 para teste, totalizando 302 dados.

3.3. Construção e Avaliação das Carteiras

A estratégia para construção das carteiras de ações proposta neste trabalho consiste em: fazer a predição do desempenho de cada uma das ações que compõem o índice usando os modelos gerados pelos algoritmos selecionados; ordenar as ações em ordem decrescente de desempenho; selecionar as 15 ações com melhor desempenho para compor a carteira sendo construída; calcular a média aritmética das ações do conjunto selecionado para obter o valor do desempenho anual da carteira para cada modelo. Foram construídas oito carteiras para o primeiro cenário, começando com os valores preditos para o ano 2014 e terminando em 2021. Já para o segundo cenário foram construídas 6 carteiras, com início no ano 2016 e fim em 2021.

Os desempenhos obtidos com as carteiras criadas com o uso dos modelos de regressão e com o SIF foram também comparados com um método manual de construção de carteira proposto por [Graham 2016] e com o índice S&P500 Ajustado, descrito na seção 3.3. O método manual proposto em [Graham 2016] consiste em: multiplicar os valores dos indicadores P/L e P/VP; ordenar as ações em ordem crescente pelo resultado da multiplicação; selecionar as 15 primeiras ações, ou seja, as com menor valor na relação $P/L * P/VP$; calcular a média aritmética do desempenho dessas 15 ações, para obter o desempenho anual da carteira manual.

Os modelos de regressão são avaliados pelo cálculo do erro do modelo, sendo a métrica utilizada neste estudo a do Erro Quadrátio Médio (EQM), definido na equação 1.

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{1}$$

onde x_i é o i-ésimo valor real, y_i é o i-ésimo valor predito e n é o número de dados utilizados.

4. Experimentos e Análises dos Resultados

Nesta seção são descritos os experimentos realizados para avaliar o desempenho dos modelos de predição e a eficácia da estratégia proposta para a construção das carteiras de

investimentos. Para a execução dos algoritmos de regressão foi utilizada a biblioteca Scikit-learn [1], e, para a geração do modelo de regras fuzzy, a ferramenta Weka [2].

Os modelos de regressão foram avaliados pelo índice EQM, calculado tanto para os dados de teste quanto para os dados de treinamento, visando verificar se os modelos sofreram sobreajuste. Para a avaliação com dados de treinamento foi utilizada validação cruzada com 5 pastas, com divisão aleatória. Para melhor análise dos resultados, os métodos de regressão foram organizados em um ranking de acordo com o valor de EQM a cada ano. As médias aritméticas das colocações dos modelos ao longo dos anos foram calculadas, gerando um ranking único para os algoritmos, de acordo com o valor de EQM.

4.1. Cenário 1

A Tabela 1 mostra os valores dos EQM dos modelos RL, RF, KNN, BR e AR dos anos de 2014 a 2021. O primeiro valor de cada coluna é o EQM da predição com os dados de teste (ano seguinte) e o segundo valor é o EQM da predição com dados de treinamento usando validação cruzada. Como esperado, os erros de predição com dados de treinamento são menores que os de dados de teste. Entretanto, com base em uma avaliação empírica, é possível afirmar que, de forma geral, a diferença entre os dois valores indica que não houve sobreajuste no treinamento. Vale destacar que, para o cenário 1, não foi implementado o modelo fuzzy.

Tabela 1. EQM das predições com dados de treinamento e teste - cenário 1

	2014	2015	2016	2017	2018	2019	2020	2021
RL	(0.0665, 0.208)	(0.0689, 0.0422)	(0.0946, 0.0513)	(0.0571, 0.0602)	(0.1155, 0.0589)	(0.2211, 0.0401)	(0.1137, 0.0515)	(0.2604, 0.0718)
RF	(0.0842, 0.0935)	(0.0755, 0.0475)	(0.1024, 0.05)	(0.0775, 0.065)	(0.1319, 0.061)	(0.232, 0.0384)	(0.1101, 0.0505)	(0.1206, 0.0627)
KNN	(0.0856, 0.1117)	(0.0758, 0.0515)	(0.0927, 0.057)	(0.0667, 0.0687)	(0.1262, 0.0623)	(0.2214, 0.0499)	(0.1203, 0.0602)	(0.1001, 0.0893)
BR	(0.0664, 0.1885)	(0.06898, 0.0415)	(0.09232, 0.0509)	(0.05636, 0.0601)	(0.11554, 0.0562)	(0.21991, 0.0384)	(0.1142, 0.0513)	(0.24147, 0.071)
AR	(0.1565, 0.1413)	(0.1169, 0.0769)	(0.1266, 0.1006)	(0.1259, 0.1099)	(0.1801, 0.1029)	(0.2565, 0.0629)	(0.1396, 0.0833)	(0.1855, 0.1129)

Na Tabela 2 pode-se observar o ranking dos modelos a cada ano e a média de suas colocações. Analisando as médias conclui-se que o modelo BR obteve melhor colocação, seguido pela regressão linear, KNN e RF, que obtiveram colocações intermediárias. Já a AR foi o modelo que obteve a pior colocação.

Tabela 2. Ordenação dos algoritmos pelo EQM médio - cenário 1

	2014	2015	2016	2017	2018	2019	2020	2021	Média
Reg. Linear	2º	1º	3º	2º	2º	2º	2º	5º	**2.3**
Rand. Forest	4º	3º	4º	4º	4º	4º	1º	2º	**3.2**
KNN	3º	4º	2º	3º	3º	3º	4º	1º	**2.8**
Bayes Ridge	1º	2º	1º	1º	1º	1º	3º	4º	**1.75**
Arv. Regres.	5º	5º	5º	5º	5º	5º	5º	3º	**4.75**

[1]https://scikit-learn.org/stable/

[2]https://www.cs.waikato.ac.nz/ml/weka/

4.2. Cenário 2

A Tabela 3 mostra os valores dos EQM dos modelos RL, RF, KNN, BR, AR e SIF dos anos de 2016 a 2021. Oobserva-se que, assim como no cenário 1, neste cenário os modelos também não sofreram sobreajuste.

Tabela 3. EQM das predições com dados de teste e de treinamento - cenário 2

	2016	2017	2018	2019	2020	2021
RL	(0.056, 0.0789)	(0.0592, 0.0907)	(0.073, 0.0523)	(0.0912, 0.0617)	(0.0673, 0.0709)	(0.2889, 0.0785)
RF	(0.074, 0.0762)	(0.0577, 0.0761)	(0.0926, 0.0531)	(0.1204, 0.0553)	(0.0672, 0.0628)	(0.1066, 0.0788)
KNN	(0.0803, 0.0933)	(0.0713, 0.1035)	(0.0804, 0.0682)	(0.0984, 0.075)	(0.0841, 0.0867)	(0.0982, 0.1027)
BR	(0.05591, 0.0746)	(0.05756, 0.0905)	(0.07297, 0.0524)	(0.09105, 0.0615)	(0.06755, 0.0707)	(0.28148, 0.0785)
AR	(0.1537, 0.1419)	(0.1029, 0.1453)	(0.1614, 0.0886)	(0.1558, 0.0953)	(0.1014, 0.1025)	(0.1685, 0.1277)
SIF	(0.081, 0.073)	(0.06, 0.077)	(0.075, 0.084)	(0.21, 0.146)	(0.093, 0.15)	(0.064, 0.171)

Na Tabela 4 pode-se observar o ranking dos modelos e a média de suas colocações. Analisando as médias conclui-se que o modelo BR obteve melhor colocação, seguido pela RL, KNN, RF e SIF, que obtiveram colocações intermediárias, já a AR foi o modelo que obteve a pior colocação. Os rankings não sofreram modificações do cenário 1 para o 2, com os modelos ocupando as mesmas posições em ambos os cenários.

Tabela 4. Ordenação dos algoritmos pelo EQM médio - cenário 2

	2016	2017	2018	2019	2020	2021	**Média**
Reg. Linear	2º	2º	2º	2º	2º	6º	**2.6**
Rand. Forest	5º	3º	5º	4º	1º	3º	**3.5**
KNN	3º	5º	4º	3º	4º	2º	**3.5**
Bayes. Ridge	1º	1º	1º	1º	3º	5º	**1.3**
Arv. Regressão	6º	6º	6º	6º	6º	4º	**5.6**
Fuzzy	4º	4º	3º	6º	5º	1º	**3.8**

4.3. Avaliação das Carteiras de Ações

Para avaliar a estratégia de alocação de carteiras, foram geradas carteiras com as 15 ações que obtiveram o melhor desempenho predito por cada um dos modelos. O desempenho predito geral de cada carteira pode então ser comparado com o desempenho real. As carteiras criadas com o uso dos modelos de regressão foram também comparadas com um método manual de construção de carteira proposto por [Graham 2016] e com o índice S&P500 Ajustado, descrito na seção 3.3. As Tabelas 5 e 6 apresentam os resultados das carteiras para os cenários 1 e 2, respectivamente. É importante destacar que para cada ano e modelo as carteiras são distintas, ou seja, são compostas por ações diferentes. Os números de tais tabelas representam o percentual de valorização ou desvalorização (real ou predito) que a carteira obteve no ano. As duas primeiras linhas, representando respectivamente as carteiras manual e com S&P Ajustado, não têm o valor predito.

As Figuras 1 e 2 mostram a evolução dos desempenhos preditos e reais, respectivamente, para o cenário 1, com os desempenhos de cada ano, sintetizados na Tabela 5.

Tabela 5. Desempenhos (em porcentagens) reais e preditos pelas carteiras geradas - cenário 1

		2014	2015	2016	2017	2018	2019	2020	2021	Média
Manual		10,0	-1,8	29,7	18,6	-17,1	30,6	-7,2	28,3	**11,3**
S&P500_A		16,4	2,1	17,3	20,8	-6,7	32,5	12,1	27,5	**15,2**
Reg.	real	35,3	28,1	7,4	17,7	-10,0	39,3	31,9	28,5	**22,3**
Linear	predito	44,0	19,7	34,7	21,8	23,5	6,1	45,0	140,0	**41,9**
Rand.	real	20,6	33,6	27,3	25,1	4,3	44,5	36,9	35,0	**28,4**
Forest	predito	71,3	40,6	42,0	55,6	48,6	12,1	57,6	64,2	**49,0**
KNN	real	12,4	1,2	19,6	33,8	-12,5	31,8	26,9	36,1	**18,7**
	predito	63,9	42,7	38,6	50,6	46,3	14,5	55,7	39,9	**44,0**
Bayes	real	35,3	25,4	7,4	23,1	-10,0	39,4	18,3	28,5	**20,9**
Ridge	predito	43,5	18,9	31,8	21,6	23,5	4,8	41,2	131,0	**39,6**
Arv.	real	16,9	12,0	27,0	19,6	-7,2	37,5	21,9	32,8	**20,1**
Regres.	predito	108,0	76,6	58,5	83,1	85,5	36,3	79,6	94,0	**77,8**

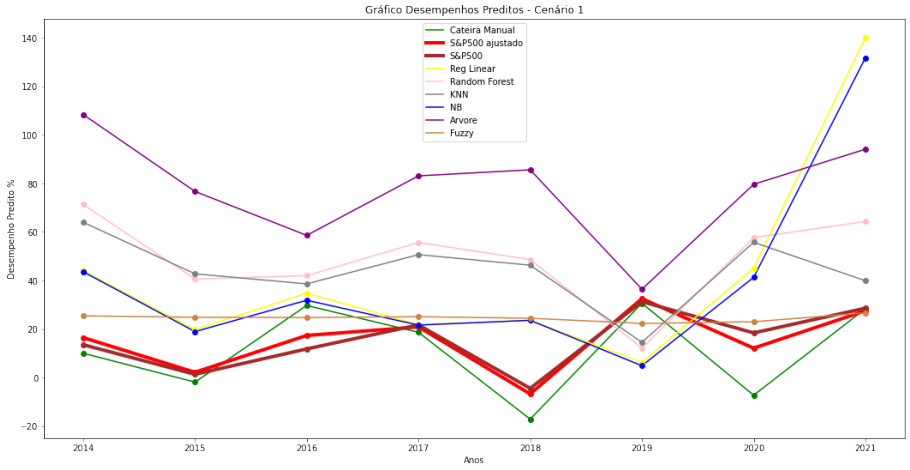

Figura 1. Desempenhos preditos pelas carteiras geradas por cada modelo.

Comparando-se a média do índice S&P500 Ajustado para o cenário 1, 15.2% com as médias dos resultados preditos, constata-se que todos os modelos de AM obtiveram desempenhos médios acima da média do índice. A árvore de regressão obteve o melhor desempenho, 77.8%, enquanto BR apresentou menor desempenho, 20.9%. Aapenas a Carteira Manual obteve um resultado abaixo do índice S&P500 Modificado, com 11.3%. Analisando a média dos desempenhos reais das carteiras montadas pelos métodos de regressão, constata-se também que todas elas obtiveram desempenhos superiores a média de mercado, com o RF apresentando melhor desempenho médio, 28.4%, e o KNN com o pior desempenho médio, 18.3%. Quando compara-se a carteira manual com os modelos de AM, observa-se que média de desempenho da carteira manual foi abaixo dos demais métodos e, em algumas raras exceções como 2017 e 2021, teve performance superior a RL e S&P500 Ajustado.

Nota-se também que a AR apresentou, em alguns anos, grande disparidade entre as médias dos valores reais e preditos, sugerindo que o modelo pode se comportar de maneira instável em determinados anos. Já o BR e RL apresentaram menor diferença

Figura 2. Desempenhos reais pelas carteiras geradas por cada modelo.

entre os valores reais e preditos, mostrando também uma baixa variância nos valores reais ao longo dos anos, com resultado abaixo do índice de mercado apenas no ano de 2016 e 2018 para o BR e 2017, 2018, 2019 para a RL. Com isso, constata-se que os modelos apresentaram estabilidade, como também desempenho constante eacima da média.

Tabela 6. Desempenhos (em porcentagens) reais e preditos pelas carteiras geradas - cenário 2

		2016	2017	2018	2019	2020	2021	Média
Manual		37,3	17,0	20,1	22,0	22,8	5,2	**20,7**
S&P500_A		17,3	20,8	-6,7	32,5	12,1	27,5	**17,2**
Reg.	real	27,5	26,2	3,8	44,0	41,2	20,6	**27,2**
Linear	predito	24,5	32,9	26,3	23,9	28,4	154,4	**48,4**
Rand.	real	10,0	40,4	-0,19	41,7	51,8	19,0	**27,3**
Forest	predito	54,1	56,0	43,2	45,9	60,0	63,0	**53,7**
KNN	real	20,9	16,5	-9,1	43,4	20,0	26,3	**19,7**
	predito	67,1	46,6	35,9	48,3	33,4	43,0	**45,7**
Bayes	real	28,2	26,5	3,8	44,0	40,1	20,6	**27,2**
Ridge	predito	24,3	29,3	25,6	23,4	27,4	151,1	**46,8**
Arv.	real	15,4	39,5	-5,6	38,6	41,3	30,7	**26,6**
Regres.	predito	103,4	97,9	88,3	66,3	77,3	80,9	**85,7**
Fuzzy	real	2,7	30,8	-6,3	37,5	29,7	37,5	**22,0**
	predito	50,0	50,0	50,0	50,0	50,0	50,0	**50,0**

Para o cenário 2 é importante frisar que há o acréscimo da carteira gerada pelo SIF nas análises. Comparando-se a média do índice S&P500 Ajustado para o cenário 2, 17.2%, com as médias dos resultados preditos, constata-se que todos os modelos de regressão obtiveram desempenhos acima da média do índice, com a AR obtendo o melhor desempenho, 85.7%, enquanto KNN apresentou menor desempenho, 45.7%. Analisando-se os desempenhos médios reais, constata-se também que todas as carteiras geradas pelos modelos de regressão obtiveram desempenhos superiores a média de mercado, com o RF, RL e BR apresentando desempenhos muito próximos. Já o SIF apresentou certa volatilidade no decorrer dos anos, mas obteve também um média de desempenho real acima

do índice S&P500 Ajustado. Quando compara-se a carteira manual com os modelos de AM, observa-se que média de desempenho da carteira manual foi maior que o S&P500 Ajustado e KNN, obtendo uma performance melhor que a obtida no cenário 1.

Assim como no cenário 1, a AR apresentou, em alguns anos, grande disparidade entre as médias dos valores reais e preditos, sugerindo que o modelo pode se comportar de maneira instável em determinados anos. O BR e RL apresentaram menor diferença entre os valores reais e preditos, mostrando também uma baixa variância nos valores reais ao longo dos anos, com resultado abaixo do índice de mercado apenas no ano 2021. Com isso constata-se que os modelos apresentaram estabilidade, como também desempenho constante e acima da média.

As Figuras 3 e 4 mostram a evolução dos desempenhos preditos e reais, respectivamente, para o cenário 2, com os desempenhos de cada ano, sintetizados na Tabela 6.

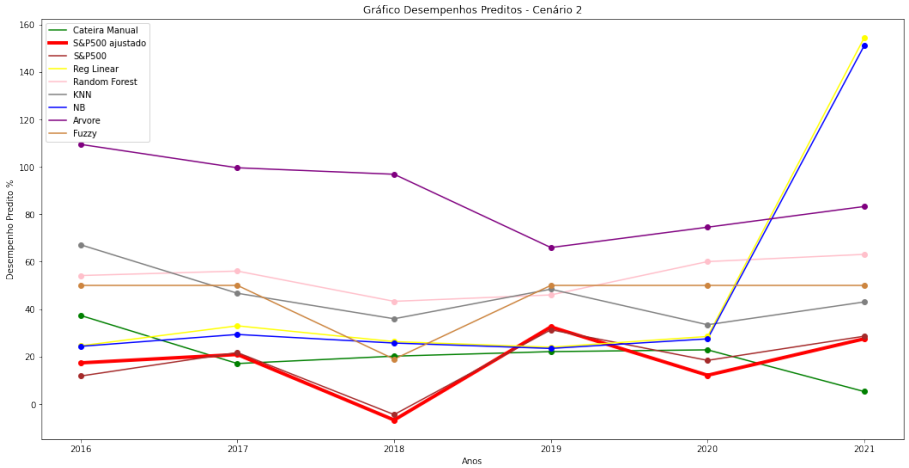

Figura 3. Desempenhos preditos pelas carteiras geradas - cenário 2.

Figura 4. Desempenhos reais pelas carteiras geradas - cenário 2.

5. Conclusões e Trabalhos Futuros

Neste trabalho, modelos de AM foram utilizados para prever os valores de ações e construir carteiras de investimentos que pudessem ter desempenho mais altos do que as carteiras montadas de forma manual e do que o próprio índice de mercado. Os resultados mostraram que a estratégia de utilização de AM para alocação de carteiras permite que o investidor selecione ações com grande potencial de valorização, com performances melhores do que o índice S&P500 Ajustado e metodologias tradicionais para construção de carteiras manuais.

Para dar continuidade à esta pesquisa, pretendemos incluir experimentos com dados referentes a um perído maior que dez anos e com um maior número de indicadores como valores de Dividendos anuais, Ebitda, ROE, dentre outros, que podem trazer mais informações sobre as empresas e consequentemente uma maior assertividade para os modelos. Outras possibilidades de trabalhos futuros incluem alteração de parâmetros do sistema fuzzy como número e formato de conjuntos fuzzy, e funções de agregação generalizadas para o cálculo das inferências.

Referências

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, 32(3):663–682.

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of financial economics*, 12(1):129–156.

Bisht, K. and Kumar, A. (2022). Stock portfolio selection hybridizing fuzzy base-criterion method and evidence theory in triangular fuzzy environmennt. *Oper. Res. Forum*, 3(53):1–33.

Castro, S. R. d. C. (2009). *Alocação de carteiras de ações através da utilização de modelos de lógica Fuzzy*. PhD thesis.

Das, R., S., S., and Maulik, U. (2022). A survey on fuzzy deep neural networks. *ACM Computing Surveys*, 53(3):54.1–54.25.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). Inteligência artificial: uma abordagem de aprendizado de máquina.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Franchi, B. O. (2021). Análise comparativa das metodologias de markowitz, kelly e aprendizado por reforço em carteiras de investimentos-uma abordagem computacional. Trabalho de conclusão de curso, Universidade Federal de São Paulo. https://repositorio.unifesp.br/handle/11600/60396.

Gambim, M. L. (2022). Alocação de carteiras de ações utilizando aprendizado de máquina e regras fuzzy. Trabalho de conclusão de curso (engenharia de computação), Universidade Federal de São Carlos. https://repositorio.ufscar.br/handle/ufscar/16715.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.".

Graham, B. (2016). *O investidor inteligente*. HarperCollins Brasil.

Han, J., Kamber, M., and Pei, J. (2012). 8 - classification: Basic concepts. In Han, J., Kamber, M., and Pei, J., editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 327–391. Morgan Kaufmann, Boston, third edition edition.

Hao, A., H., Z., and Y., Z. (2023). Stock protfolio mangement by using fuzzy ensemble deep reinforcement learning algorithm. *Journal of Risk and Financial Management*, pages 1–14.

Indices, S. D. J. (2016). S&p dow jones indices. *Retrieved April*, 12:2016.

Jiménez-Preciado, A. L., Venegas-Martínez, F., and Ramírez-García, A. (2022). Stock portfolio optimization with competitive advantages (moat): A machine learning approach. *Mathematics (Basel)*, 10(23):4449.

Klir, G. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.

MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, 4(3):415–447.

Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13.

Mazraeh, A. B., Daneshvar, A., Zaj, M. M., and Roodposhti, F. R. (2022). Stock portfolio optimization using a combined approach of multi objective grey wolf optimizer and machine learning preselection methods. *Computational Intelligence and Neuroscience*, pages 1–20.

Mendel, J. M. and Bonissone, P. P. (2021). Critical thinking about explainable ai (xai) for rule-based fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 29(12):3579–3593.

Orang, O., de Lima e Silva, P. C., and Guimarães, F. G. (2022). Time series forecasting using fuzzy cognitive maps: a survey. *The Artificial intelligence review*, 56(8):7733–7794.

Ross, S. A. (2009). Neoclassical finance. In *Neoclassical Finance*. Princeton University Press.

Santos Junior, J. G. A. (2015). Um estudo sobre aprendizado de máquina aplicado à modelagem de retornos de ações. Dissertação de mestrado, UFRN. https://repositorio.ufrn.br/jspui/handle/123456789/26066.

Skrjanc, I., Iglesisas, J. A., Sanchis, A., Leite, D., Lughofer, E., and Gomide, F. (2019). Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification and classification: A survey. *Information Siences*, 49:344–368.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

Zadeh, L. (1965). Fuzzy sets. *Information and control*, pages 338–338.