Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Trabalho de Conclusão de Curso em Engenharia de Computação–

Mario Luiz Gambim

Alocação de carteiras de ações utilizando aprendizado de máquina e regras Fuzzy

São Carlos
2022

Mario Luiz Gambim

Alocação de carteiras de ações utilizando aprendizado de máquina e regras Fuzzy

Monografia apresentada como Trabalho de Conclusão de Curso em Engenharia de Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de bacharel em Engenharia de computação.

Orientador: Heloísa de Arruda Camargo

São Carlos
2022

Dedico este trabalho a meus pais Luiz Fernando Gambim e Mary Delforno Gambim que acreditaram na educação como instrumento de mudança.

Agradecimentos

Aos meus pais e meu irmão, Luiz Fernando, Mary e Gustavo, que sempre estiveram ao meu lado dando todo suporte necessário para que eu pudesse chegar até aqui. A vocês dedico todo amor.

A minha namorada, Giuliana Balieiro Crescitelli, que esteve ao meu lado dando apoio incondicional durante a vida acadêmica e sempre acreditou no meu potencial.

A minha família, em especial a meus primos Danielle e Lucas Petto que sempre se mantiveram presentes durante toda minha vida estudantil, dando suporte e incentivo.

Aos meus amigos da graduação, que me ajudaram nos momentos mais difíceis de toda caminhada, sou eternamente grato.

A minha orientadora, Heloísa de Arruda Camargo, que acreditou no tema proposto e sempre se mostrou extremamente dedicada a me auxiliar ao máximo no trabalho.

"Tudo vale a pena quando a alma não é pequena."
(Fernando Pessoa)

Resumo

Os investimentos de longo prazo por meio de carteiras de ações atrai a atenção de investidores, que buscam formas simples e eficazes de selecionar ações para compor uma carteira, assim como encontrar uma série de ativos daqueles que terão maior valorização no futuro. Ao longo dos anos várias técnicas foram desenvolvidas, e esse trabalho teve como objetivo a utilização de Aprendizado de máquina (AM) para previsão dos valores futuros de ações.

Para maximizar seus ganhos os investidores buscam por informações ou técnicas que sejam capazes de auxilia-los, diante disso, modelos de AM passaram a ser frequentemente utilizados para descoberta de ações que tenham potencial de valorização elevado. Ao contrário de aplicações de curto e médio prazo que fazem uso de séries temporais, as aplicações de longo prazo requerem métodos mais simples.

Neste trabalho foi elaborada e avaliada uma estratégia para alocação de carteiras para aplicações de longo prazo (1 ano) com modelos de AM, são eles: Regressão Linear (RL), Árvore de Regressão (AR), Random Forest (RF), K-Nearest Neighbors (KNN), Bayesian Ridge (BR) e Sistemas de Inferência Fuzzy (SIF). Em complementação a esse objetivo principal, foi feita também uma análise qualitativa da interpretabilidade das regras de um modelo baseado em regras fuzzy, para buscar meios de compreender a relação entre atributos de entrada e de saída com base em termos linguísticos.

O resultado obtido neste trabalho mostrou que a abordagem pelo uso de modelos de AM para predição do desempenho das ações ao longo de um ano e posterior seleção das ações é eficiente para investidores obterem rendimentos superiores aos índices de mercado S&P500 e por métodos manuais.

Palavras-chave: aprendizado de máquina, carteira de investimento, regressão, sistemas fuzzy.

Abstract

Long-term investments through equity portfolios attract the attention of investors, who seek simple and select stocks to compose a Portfolio, as well as finding a series of assets that will have greater dignity no future. Over the years, several techniques have been developed, and this work had as an objective the use of Machine Learning to predict the values stock futures.

To maximize their earns, investors look for information or techniques that addition power to assist, on, AM models them to be requested for the discovery of used shares with high appreciation potential. To alternatives of short and medium term applications that make use of time series, the Long-term prescribing simpler methods.

In this work, a survey was carried out and one for the allocation of portfolios to long term applications (1 year) with Machine Learning models, they are: Linear Regression, Regression Tree, Random Forest, K-Nearest Neighbors, Bayesian Ridge and Fuzzy Inference Systems. In addition to this objective principal, a qualitative analysis of the interpretability of the analysis rules was also carried out a model based on fuzzy rules search, for means of understanding the relationship between input and output attributes based on linguistic terms.

The result obtained in this work showed that the approach by using models of Machine Learning to predict the performance of stocks over a year and subsequent selection of shares is efficient for investors to obtain returns above market indices S&P500 and by methods.

Keywords: machine learning, stock investiment, regression, fuzzy system.

Lista de ilustrações

Lista de tabelas

Lista de siglas

P/L Preço/Lucro

P/VP Preço/Valor Patrimonial

VOL Volume

IA Inteligência Artificial

AM Aprendizado de máquina

EQM Erro Quadrático Médio

EAM Erro Absoluto Médio

RL Regressão Linear

AR Árvore de Regressão

VC Validação Cruzada

RF Random Forest

KNN K-Nearest Neighbors

BR Bayesian Ridge

SIF Sistemas de Inferência Fuzzy

Sumário

Capítulo 1

Introdução

A constante evolução do mercado financeiro ao longo dos anos tornou esse segmento mais robusto e complexo, no que diz respeito as operações envolvidas como também gerou uma maior gama tipos de aplicações, podendo elas serem de renda fixa e 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 investimento, que pode ser de curto, médio ou longo prazo e para isso buscam informações ou métodos que sejam capazes de apontar as melhores aplicações. Esse trabalho irá utilizar estudos que identificaram múltiplos de mercado capazes de explicar o retorno futuro das ações Basu (1983), contrariando a teoria dos mercados eficientes, na qual diz que qualquer informação disponível é instantaneamente assimilada pelos operadores de mercado, impedindo que qualquer investidor obtenha vantagem na obtenção de lucros (ROSS, 2009). Diante disso, métodos computacionais passaram a ser aplicados por investidores com o objetivo de encontrar as melhores opções. Dentre eles podemos destacar métodos de 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 futuro, também o trabalho de Franchi (2021) que fez uso de aprendizado por reforço para otimização de carteiras de investimento e Junior (2015) que utilizou máquinas de vetores de suporte para predição dos valores dos preços de ações.

Entre os diversos tipos de aplicação, as de longo prazo estão entre os interesses de investidores. Para esse tipo de aplicação, o investidor deseja selecionar ações de empresas para construir uma carteira de investimentos e buscam informações de diversas fontes incluindo o 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. O uso de séries temporais pode levar a um resultado

abaixo do esperado se a série for não linear contaminada por ruído, ou até se sua dinâmica for diferente do modelo auto regressivo (JUNIOR, 2015).

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 e analisar quão eficientes os modelos de aprendizado de máquina podem ser ao procurar por ações com preço de mercado subvalorizado, ou seja, buscaremos encontrar dentre um conjunto de ações as que sejam mais promissoras e eventualmente retornem um desempenho maior que a média do conjunto estudado. Em complementação a esse objetivo principal, foi feita também uma análise qualitativa da interpretabilidade das regras de um modelo baseado em regras fuzzy, para buscar meios de compreender a relação entre atributos de entrada e de saída com base em termos linguísticos.

A estratégia proposta nesse trabalho é aplicada por meio de uma análise fundamentalista apresentada por Graham (2016), que determina que os múltiplos P/L e P/VP 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 pelo seu lucro total, ou seja, ações com baixo P/L indicam que uma empresa tem um alto lucro frente ao valor de suas ações. Já o P/VP é a razão entre o preço das ações e o valor patrimonial da empresa, ou seja, o preço das ações frente ao seu patrimonio, e um baixo P/VP também pode indicar que a ação daquela empresa apresenta valores descontados. Além dos desses dois múltiplos foi adicionado o atributo volume de negociação, que para esse trabalho é 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). A partir desses valores de desempenho, as ações com maior desempenho predito são selecionadas para compor a carteira. Foram avaliados os desempenhos dos modelos comparando resultados das predições com dados de treinamento e teste. Depois foram avaliadas as carteiras construídas usando os modelos e comparadas com a carteira manual proposta por Graham (2016) e com S&P500 Ajustado. Posteriormente, foi criada e avaliada carteiras de ações contendo as 15 ações com melhor desempenho predito por cada modelo, e novamente avaliadas.

Primeiro foram avaliados os desempenhos dos modelos comparando resultados das predições com dados de treinamento e teste. Depois foram avaliadas as carteiras construídas

usando os modelos e comparadas com a carteira manual e com S&P500.

Os resultados demonstraram que o uso de modelos de AM para predizer o desempenho das ações no fechamento do ano e a seleção das ações com maior desempenho predito leva à alocação de carteiras com desempenho superior às das carteiras criadas por métodos manuais ou por índices tradicionais de mercado.

Este trabalho está organizado da seguinte forma:

- Capítulo 2: apresentada a fundamentação teórica, com os conceitos necessários para compreensão dos assuntos mencionados no decorrer do trabalho e é realizada a apresentação dos trabalhos relacionados, com abordagens semelhantes ao deste.
- Capítulo 3: é apresentado a geração dos modelos de regressão utilizados, a construção dos SIF e a estratégia para alocação de carteiras proposta e analisada neste trabalho.
- Capítulo 4: são descritos os experimentos e análise dos resultados, através da avaliação dos modelos, e apresentação das carteiras.
- Capítulo 5: apresenta as conclusões e trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo serão apresentados conceitos básicos necessários para compreensão do trabalho, tanto da área de alocação de carteiras de ações quanto da área de aprendizado de máquina. Também serão demonstrados trabalhos que abordam problemas semelhantes ao dessa monografia.

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. Diante disso, investidores procuram formular suas estratégias de investimentos baseados nessas teorias, para que assim possam maximizar seus ganhos e minimizar seus prejuízos. Duas dessas estratégias baseiam-se na Hipótese dos Mercados Eficientes e na Análise Fundamentalista que serão abordadas nos tópicos posteriores.

2.1.1 Hipótese dos Mercados Eficientes

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, e assumindo que preço (do ativo no momento da compra) e valor, que pode ser definido pelas caracterís-

ticas da corporação representada por esse ativo, como governança, endividamento, setor de atuação, dentre outras de um ativo sejam iguais. No entanto, Hipótese dos Mercados Eficientes também considera que eventualmente os dois possam assumir valores diferentes, no entanto 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.

2.1.2 Análise Fundamentalista

A 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.

Atualmente há uma gama de indicadores fundamentalistas que servem de base para os investidores tirarem conclusões sobre empresas e formular a melhor tese de investimentos, dentre esses indicadores um dos mais utilizados é o P/L. O preço por ação é o preço de mercado de uma ação, enquanto o lucro por ação é o lucro líquido mais atual referente ao período de um ano, dividido pelo número de ações existentes para aquele empresa. Estudos de Basu (1977) que 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 a análise de uma ação não é eficiente baseando-se somente nesse múltiplo, uma vez que outras características como valor de mercado da empresa e setor de atuação tem ligação direta com a razão P/L (CASTRO, 2009).

Um outro indicador fundamentalista amplamente utilizado é o P/VP que é a razão entre o preço de uma ação e o valor patrimonial da empresa. Estudos mostraram que ações negociadas a preços menores do que o valor patrimonial da empresa são subvalorizadas, em contrapartida as que apresentam preços maiores são supervalorizadas, podendo sugerir ao investidor que considere empresas com P/VP baixos no momento de montar sua carteira de ações (CASTRO, 2009).

2.2 Inteligência artificial

A citação a seguir nos da uma boa definição inicial de Inteligência Artificial (IA):

> Inteligência artificial (IA) pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente. (LUGER, 2004)

Sendo assim, segundo Luger (2004), a IA é uma área da ciência da computação que tem sua fundamentação em sólidos princípios teóricos e práticos. Esse conjunto de fundamentos são resultantes da busca do ser humano pelo maior entendimento de como ele mesmo formula suas ações e pensamentos. Para isso são usadas ferramentas computacionais poderosas como algoritmos, estruturas de dados, linguagens de programação, para replicar lógicas do pensamento humano em ambientes computacionais.

Nesse contexto, a introdução da ideia de IA como conhecemos hoje pode ser encontrada no teste de Turing, proposto pelo próprio Turing e Haugeland (1950) em que consiste em testar a capacidade de uma máquina em exibir o comportamento inteligente correspondente a um ser humano, e para este teste seria necessário o computador ter algumas capacidades, segundo Russell (2010) são elas:

- ❏ Processamento de linguagem natural para viabilizar a comunicação ;
- ❏ Representação de conhecimento para armazenamento de informações;
- ❏ Raciocínio automatizado para usar as informações com objetivo de tirar conclusões e responder a novas perguntas;
- ❏ Aprendizado de máquina para identificar padrões e adaptação a novas circunstâncias.

Para uma versão completa do teste, o computador precisaria de algumas capacidades finais:

- ❏ Visão computacional para descobrir objetos;
- ❏ Robótica para movimentar-se e mexer objetos.

Os itens citados acima já são temas de estudos para diversas áreas da IA, no entanto o desenvolvimento da IA passou por altos e baixos desde o teste proposto por Turing até os dias atuais. Os momentos de baixa devem-se, entre outras coisas, a limitações como falta de processamento computacional suficiente para resolução de problemas, como também pouca capacidade de armazenamento e de dados disponíveis para para que os algoritmos tivessem desempenho satisfatório.

Nos últimos anos, com a capacidade de processamento e de armazenamento aumentadas exponencialmente, permitiu-se que os estudos na área de IA pudessem crescer vertiginosamente e as soluções propostas pela área pudessem estar cada vez mais presentes nas empresas e na vida das pessoas.

Atualmente, dentre as diversas áreas de IA, a de AM é a área com maior atividade, sendo assim, por ser de interesse desse trabalho serão apresentados nas seções seguintes seus principais conceitos.

2.3 Aprendizado de máquina

A capacidade de aprender deve fazer parte de qualquer sistema que possui algum grau de inteligência, dessa maneira, agentes inteligentes devem ter aptidão de realizar modificações ao longo do curso de suas interações com o ambiente externo, através de experiência passadas e presentes. Sendo assim, o aprendizado abrange a propagação através da experiência, ou seja, tarefas de mesmo domínio devem servir de exemplo para a melhora no desempenho, e não apenas a repetição da mesma tarefa (LUGER, 2004).

Nesse contexto, AM se apresenta como uma área muito vasta de pesquisa. A produção de algoritmos diferentes para soluções diversas, variam de acordo com o problema em questão, no que envolve a disponibilidade de dados para treinamento, e nas estratégias de linguagem de representação do conhecimento. Assim, mesmo com diferentes abordagem, o objetivo é o mesmo, encontrar uma generalização aceitável para determinado problema (RUSSELL, 2010).

Nas seções posteriores irá se explorar melhor os temas de aprendizado supervisionado e não supervisionado, que fazem uso de conjuntos de dados estruturados na forma de tabela Objeto X Atributo.

2.3.1 Aprendizado não supervisionado

O AM não supervisionado possui como principal característica a não supervisão externa na forma de um atributo especial chamado atributo meta, que define os rótulos dos dados, ou seja, os dados não são rotulados, sendo assim, algoritmos visam encontrar estruturas intrínsecas aos dados e construir representações dessas estruturas (FACELI et al., 2011). Para exemplificar de maneira mais didática podemos associar esse tipo de aprendizado ao método científico: propor hipóteses para explicar observações, avaliar essas hipóteses usando critérios e as testar através de experimentos (RUSSELL, 2010).

Por mais que não seja fornecido nenhum retorno explícito o agente identifica padrões nos dados de entrada. Sendo assim, a tarefa mais comum para o AM não supervisionado é o agrupamento, ou seja, a identificação de grupos nos dados em função da similaridade dos objetos. Um exemplo cotidiano desse fenômeno é de um motorista que desenvolve a

ideia de "dia de tráfego bom"ou "dia de tráfego ruim"por percepções próprias, isto é, sem que nenhum professor tenha dado esse rótulo anteriormente (LUGER, 2004).

2.3.2 Aprendizado supervisionado

O AM supervisionado diferentemente do não supervisionado utiliza dados que têm o atributo meta, e assim são rotulados. O agente usa a base de dados rotulados como pares de entrada e saída, em que a entrada é formada por um ou mais atributos de um objeto e a saída é o seu rótulo, para o treinamento de um modelo, apresentando esses dados a um algoritmo que possa aprender uma função que faça o mapeamento da entrada e saída. Dessa maneira, através de um espaço de hipóteses possíveis o algoritmo buscará por aquelas que tenham melhor desempenho. Para avaliar a precisão da hipótese, fornece-se um conjunto de dados de teste que seja distinto do conjunto usado para treinamento e conclui-se que a hipótese generaliza bem o problema se consegue prever acertadamente novos valores dados novos exemplos (RUSSELL, 2010).

Quando o rótulo dos objetos em questão for nominal, isto é, possuir valores categóricos conhecidos e em pequeno número, como cor, nível de instrução classificação. Já quando o atributo meta for um atributo numérico, como velocidade ou temperatura, temos um problema de regressão.

Neste trabalho o problema tratado é de regressão e portanto nas próximas seções serão apresentados os algoritmos de AM usados no trabalho.

2.4 Regressão

A regressão é um tipo de AM supervisionado que, diferentemente da classificação, quando é dado um conjunto de dados de treinamento a função prevê valores contínuos. Dentro de regressão existem uma vasta quantidade de algoritmos que se encaixam melhor para cada problema proposto, levando em conta os tipos, organização e estrutura do conjunto de dados.

2.4.1 Regressão Linear

Um dos mais difundidos modelos de regressão são os modelos lineares, em que dado um conjunto de dados espera-se que o valor destino seja a combinação linear dos atributos de entrada. Quando temos vários atributos em nosso conjunto de dados, chamamos de regressão linear múltipla. Nessa situação há algoritmos que lidam de forma satisfatória com esses problemas, é o caso do método dos mínimos quadrados, a regressão logística dentre outros (PEDREGOSA et al., 2011).

Na Figura 1 a seguir observa-se o exemplo de uma regressão linear múltipla, na qual o plano azul é a regressão linear em que são analisadas três variáveis, sendo uma dependente

e é a função de outras duas que se comportam de maneira independente, já os pontos amarelos são os dados observados.

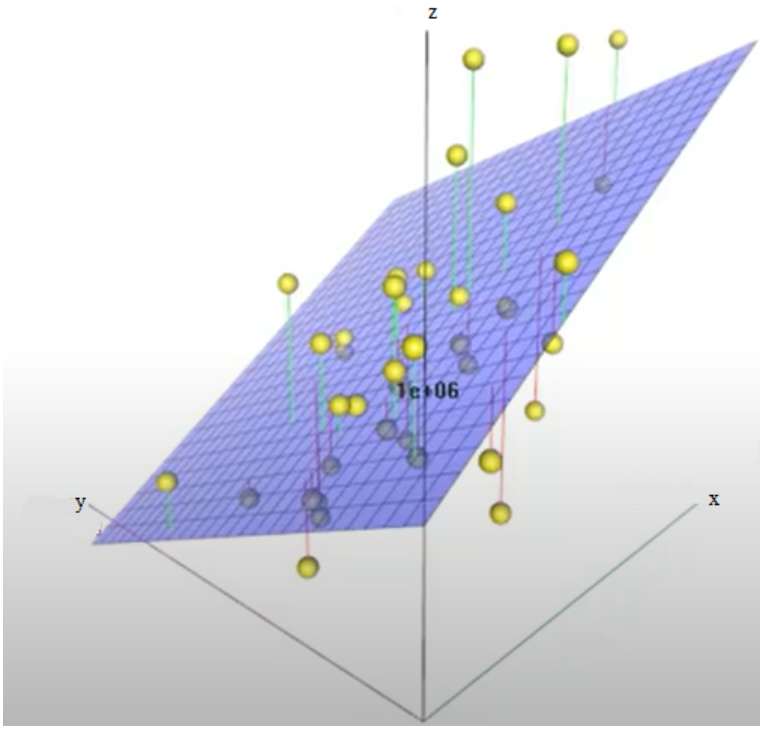

Figura 1 – Regressão Linear Múltipla.

2.4.2 Árvores de decisão e regressão

Outros métodos difundidos de AM supervisionado são os de árvores de decisão e regressão, que são que modelos partem da premissa de dividir para conquistar para resolver um problema. Um problema complexo é dividido em problemas menores, ao passo que recursivamente é repetida a mesma ideia, em que as soluções dos problemas menores podem ser usadas para produzir a solução do problema mais complexo. 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).

Na Figura 2 observa-se a representação de uma árvore de decisão, elas são bastante intuitivas e eficientes, com sua construção começando pela raiz e a cada nó criado um atributo é selecionado para esse nó, e a seleção desse podendo ser feita por diferentes índices, como ganho de informação, gini, dentre outros, assim cada nó folha terá uma classe para problemas de classificação e um valor numérico ou equação para problemas de regressão.

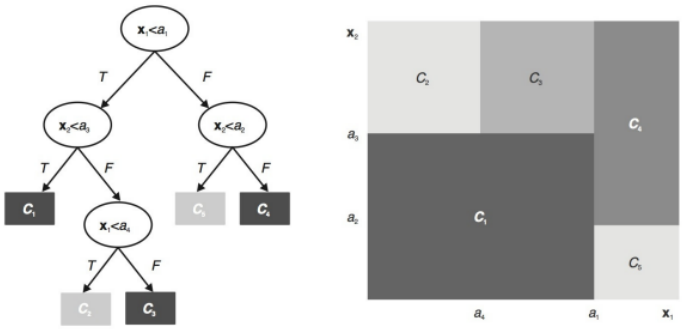

Figura 2 – Uma árvore de decisão e as regiões de decisão no espaço de objetos.

Fonte: (FACELI et al., 2011)

As árvores de decisão como já diz o nome, tem como entrada um conjunto de valores que podem ser discretos ou contínuos e retorna uma saída, a "decisão". Como dito anteriormente uma árvore alcança seu objetivo executando uma sequência de teste recursivamente, em que a cada nó interno da árvore corresponde a um teste do valor de um dos atributos de entrada, e suas ramificações são os possíveis valores dos atributos. Por fim, cada nó folha da árvore especifica o valor a ser retornado pelo algoritmo (RUSSELL, 2010).

2.4.3 Random Forest

O modelo Random Forest (RF) é uma generalização das árvores de decisão tradicionais, pode ser utilizado para classificação ou regressão e 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 serem mais rápidos e tem como principal objetivo minimizar o sobre-ajuste, que é quando uma função se ajusta muito bem ao conjunto de dados observado e mostra-se ineficaz para prever novos resultados. (HAN; KAMBER; PEI, 2012).

Como mencionado por Han, Kamber e Pei (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 classificar uma nova instância de dado, cada árvore é percorrida para definir sua saída e a classe final é definida pela maioria de votos de todas as árvores.

2.4.4 K-Nearest Neighbors

O K-Nearest Neighbors (KNN) é um método de AM que é baseado em aprendizado por comparação, ou seja, tuplas de teste são comparadas a tuplas de treino semelhantes

a ela. É um algoritmo de aprendizado lazy learning, no qual não é feita a generalização do conhecimento e o modelo é formado pelas próprias instâncias, que são armazenadas. As tuplas usadas para treinamento podem conter n atributos e cada uma representa um ponto em um espaço de *n-dimensões*, e são armazenadas num espaço de padrão também *n-dimensional*. 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 essa ser a euclidiana e a classificação se dará de acordo com a classe da maioria das tuplas mais próximas, ou seja, se para um k igual a 3, tivermos as duas menores distância para a classe "A"e uma terceira menor distância para a classe "B", por voto vencido a tupla será classificada como "A". Quando o problema é de regressão a saída prevista se dá pela média ou na mediana dos valores de saída das instâncias mais semelhantes (HAN; KAMBER; PEI, 2012).

2.4.5 Bayesian Ridge

O ridge regressor 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 é sob o ponto de vista Bayesiano, nesse trabalho será implementado o algoritmo Bayesian Ridge (BR) da biblioteca scikit-learn baseado no algoritmo descrito por Tipping (2001) e MacKay (1992).

2.5 Sistemas Fuzzy

Os conjuntos fuzzy foram apresentados inicialmente por Zadeh (1965) e a ideia era replicar a lógica do pensamento humano em lidar com processos complexos, que são baseados informações imprecisas e fronteiras mal definidas. A solução para isso foi a construção de modelos que possam ser utilizados para traduzir em termos matemáticos as informações imprecisas expressas por regras linguísticas.

2.5.1 Conjuntos Fuzzy

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. A situação imposta é que no cotidiano os conjuntos clássicos não são capazes de lidar com

determinados tipos de elementos, 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 (PIMENTA, 2009).

Inicialmente propostos por Zadeh (1965), os conjuntos fuzzy são definidos por uma função que generaliza a função característica dos conjuntos clássicos. Tal função é chamada de função de pertinência e o conjunto por ela definido é o conjunto fuzzy (KLIR; YUAN, 1995). Pode-se denotar a função de pertinência como $A : X -> [0, 1]$ em que conecta os elementos do conjunto universo X a valores reais do intervalo $[0, 1]$. Dessa maneira, torna-se possível representar pertinências parciais entre 0 e 1, sendo 0 para nenhum grau de pertinência e 1 para pertinência total (YAGUINUMA, 2013).

2.5.2 Funções de pertinência

As funções de pertinência mais comuns são a triangular, trapezoidal e gaussiana (KLIR; YUAN, 1995).

Nesse trabalho, faremos uso somente da função triangular, por isso nos limitaremos a falar somente dela.

A função triangular pode ser representada na forma (KLIR; YUAN, 1995):

$$A(x) = max(min(\frac{x-a}{m-a}, \frac{b-x}{b-m}), 0)$$

Na Figura 3 pode-se observar a representação da forma triangular do conjunto Fuzzy.

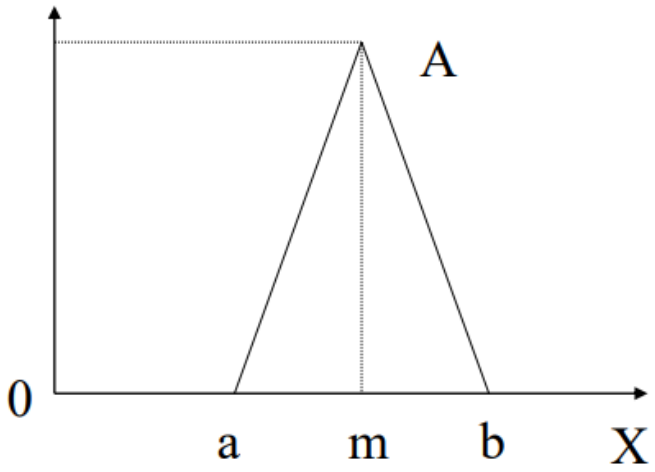

Figura 3 – Forma triangular do conjunto Fuzzy.

em que:

- m é o valor modal de μ ;
- a e b são os limites inferior e superior;
- E [a, m, b] podem ser interpretadas como as abscissas dos três vértices do triângulo.

2.5.3 Sistemas de Inferência Fuzzy

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 em aplicações em áreas de classificação de dados, controle, predição de séries temporais, tomada de decisões entre outros (JANG; SUN, 1993).

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 ao executar a inferência sobre as regras fuzzy retorna os fatos para formar a saída (PIMENTA, 2009).

Nesse trabalho utilizaremos o modelo de Mamdani e 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 resultante da inferência de cada regra. Para obter-se uma saída final não fuzzy é aplicado um método de defuzificação que será explicado no próximo tópico.

A Figura 4 ilustra o modelo de Mamdani e Assilian (1975) com três regras ($R1$, $R2$ e $R3$) e dois atributos de entrada (V_1 e V_2). Ao apresentar os valores de entrada a_1 e a_2, seus graus de pertinência são calculados nos respectivos conjuntos fuzzy (A_{i1} e A_{i2}) e combinados pelo mínimo, resultando nos graus de disparo de cada regra w_1, w_2 e w_3. Os conjuntos de saída de cada regra $B\text{t}_1$, $B\text{t}_2$ e $B\text{t}_3$ são agregados pela operação de união pelo máximo, resultando no conjunto B, que será posteriormente defuzificado para se obter uma saída numérica.

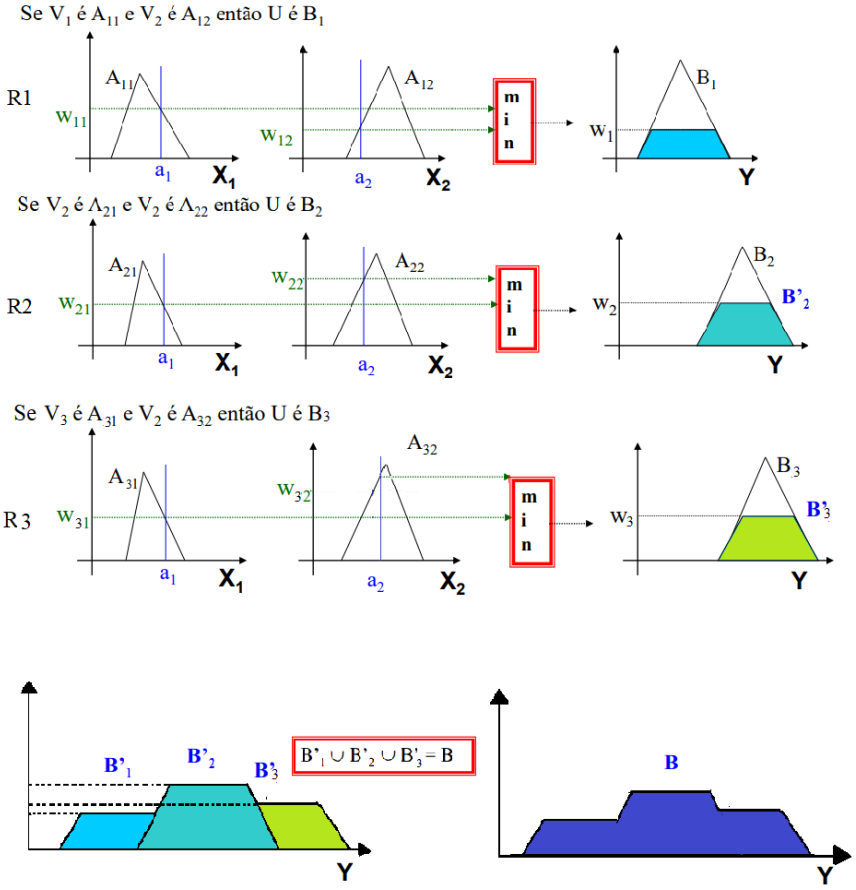

Figura 4 – Regras representadas pelos conjuntos Fuzzy.

2.5.4 Métodos de Defuzificação

Uma etapa importante nos modelos fuzzy é a interpretação e utilização dos conjuntos fuzzy que resultam dos processos de inferência, podendo ser feitas de maneiras distintas de acordo com o tipo de sistema e aplicação. No nosso caso, queremos converter a saída em valores valores numéricos, ou seja, defuzificar. Para isso alguns métodos podem ser utilizados, nesse trabalho utilizaremos o Centro de Máximos Ponderado, que pode ser descrito da seguinte maneira:

$$CoMP(B) = \frac{\sum_{R_k \in K} Vtipico(R_k) * DoF(R_k)}{\sum_{R_k \in K} DoF(R_k)}$$

em que K é o conjunto de regras disparadas (grau de disparo > 0), $DoF(R_k)$ é o grau de disparo da regra R_k e $Vtipico(R_k)$ é o valor típico do conjunto de saída da regra R_k .

2.6 Avaliação de Modelos de Regressão

Uma das maneiras mais comuns de avaliar um modelo de regressão que estima valores futuros é através do cálculo do erro do modelo. Uma vez que quanto menor o erro da saída de erro de um modelo, melhor o modelo se ajustou ao conjunto de dados e também mais próximo de valores reais a predição chegou. As métricas para cálculo do erro apresentadas serão o Erro Quadrático Médio (EQM) e o Erro Absoluto Médio (EAM).

Na Figura 5 é ilustrada uma reta de regressão, e cada ponto vermelho é a coordenada predita de um elemento no plano cartesiano, e o erro do modelo é calculado com base nas distâncias entre os pontos e a reta.

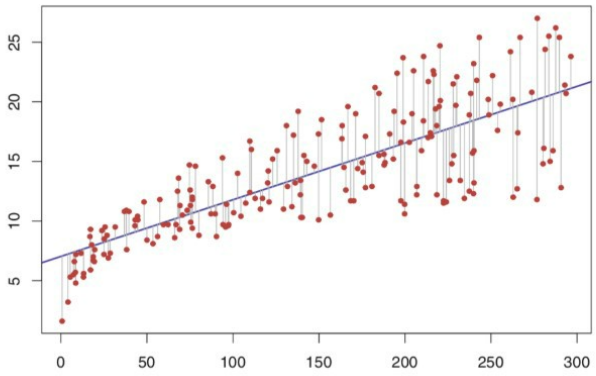

Figura 5 – Reta de regressão e desvios dos pontos analisados.

Fonte: (MELO, 2022)

2.6.1 Erro Quadrático Médio

O EQM uma das medidas de erro utilizadas para avaliar modelos preditivos, e sua equação é definida por:

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

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.

2.6.2 Erro Absoluto Médio

O EAM é outra medida de erro para avaliar modelos preditivos, e sua equação é definida por:

$$EQM = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n}$$

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.

2.7 Trabalhos relacionados

O trabalho de Castro (2009) utiliza uma abordagem interessante para tomada de decisão na compra de ações pertencentes ao índice S&P500, com o objetivo se superar o rendimento dos principais índices de mercado. Para isso foi utilizado um modelo de lógica fuzzy, onde como atributos de entrada foram usados os múltiplos Preço/Lucro e Preço/Valor Patrimonial das empresas de cada ação considerada. E para comparação da eficiência do modelo, foi implementado um modelo de regressão linear multivariada.

Já o trabalho desenvolvido por Franchi (2021) foi realizar diferentes abordagens de otimização de carteiras de investimentos propostos por Markowitz, Kelly e aprendizado por reforço. Para isso foram utilizadas séries históricas das ações presentes na Bolsa de Valores de São Paulo. Sendo assim, uma comparação foi feita entre os resultados dos retornos financeiros e computacional dos resultados obtidos, verificando o desempenho por inteligencia artificial frente a outras estratégias.

Por fim, Junior (2015) realizou um estudo sobre a aplicação de máquinas de vetores de suporte na previsão para predição nas taxas de retornos futuros dos preços das ações do mercado brasileiro, com base em valores anteriores das taxas de retorno e volatilidade. Por fim foram feitas comparações do desempenho de diversos modelos (lineares, não lineares baseados em máquinas de vetores de suporte e híbridos) em séries temporais com amostragens semanal diária e intraday de dez minutos.

Capítulo 3

Alocação de carteira de ações

Neste capítulo será descrita a estratégia proposta para alocação de carteiras de ações utilizando AM, 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.

O trabalho desenvolvido visa atender um objetivo principal, que é propor e avaliar uma estratégia de alocação de carteiras com base na predição de modelos de regressão e, complementarmente, avaliar a interpretabilidade do modelo de forma qualitativa usando as regras do modelo de inferência fuzzy.

3.1 Coleta de Dados

Para a construção do trabalho, foram utilizadas ações que compõem o índice S&P500, pertencente ao mercado americano, no ano de 2022. O índice S&P500 que 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).

Visando aumentar a fonte de informações que possam auxiliar a construção dos modelos de previsão, será considerado também, além dos múltiplos P/L e P/VP, o volume de cada ação como variável de entrada.

- **P/L**: $\frac{Preço}{Lucro}$;
- **P/VP**: $\frac{Preço}{ValorPatrimonial}$;

- **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;

Para ter acesso aos indicadores fundamentalistas usados no trabalho, P/L e P/VP foi utilizada a base de dados da Bloomberg (www.bloomberg.com), empresa que fornece informações sobre o mercado. Já para as informações referentes ao Desempenho e VOL foram retiradas do site Barchart (www.barchart.com), outra plataforma que fornece dados financeiros.

A periodicidade dos dados é anual, uma vez que queremos analisar as ações para um investimento a longo prazo, e para isso uma sazonalidade maior seria mais indicada.

O objetivo do trabalho é fazer a predição dos valores de desempenho citados acima.

3.2 Análise exploratória de dados

A análise exploratória dos dados tem como objetivo buscar entender o comportamento dos dados, como sua dispersão e correlação, sendo esses os dois pontos abordados nessa seção.

Para este trabalho foram feitos os gráficos de dispersão par a par, ou seja, um atributo em função do outro, e observou-se que os dados comportaram-se de maneira parecida, com alta concentração em determinadas regiões do plano. Como exemplo, na Figura 6 pode-se observar os gráficos de dispersão dos dados referentes ao ano de 2021, em que foram plotados os gráficos dos atributos: P/L x VOL, P/L x Desempenho, P/L x P/VP, P/VP x VOL, P/VP x Desempenho e Desempenho x VOL.

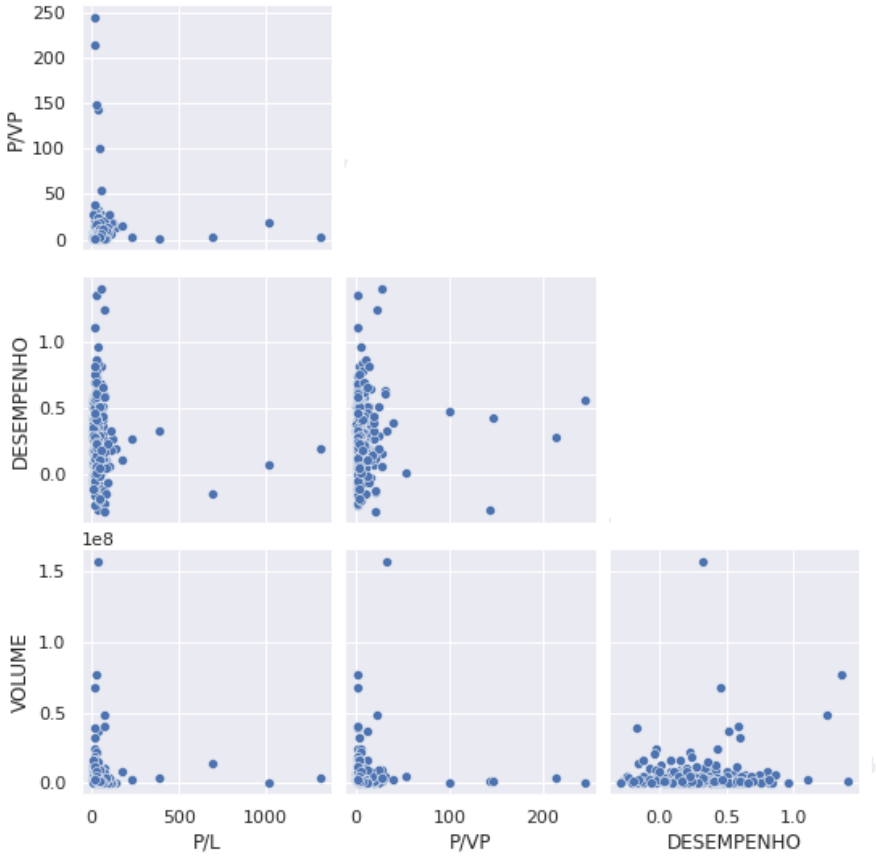

Figura 6 – Dispersão dos dados - 2021.

Já a correlação mostra o quão dependentes entre si são os atributos. Para este trabalho foi analisada a correlação entre os atributos P/L, P/VP, VOL e Desempenho. Na figura 7 observa-se as correlações dos atributos referentes aos dados do ano 2021, como os valores das correlações foram próximos de 0 pode-se concluir que a correlação entre eles é fraca (MUKAKA, 2012). Uma vez que não existe correlação significativa entre os atributos, esse fator não irá influenciar nos resultados dos algoritmos implementados no trabalho.

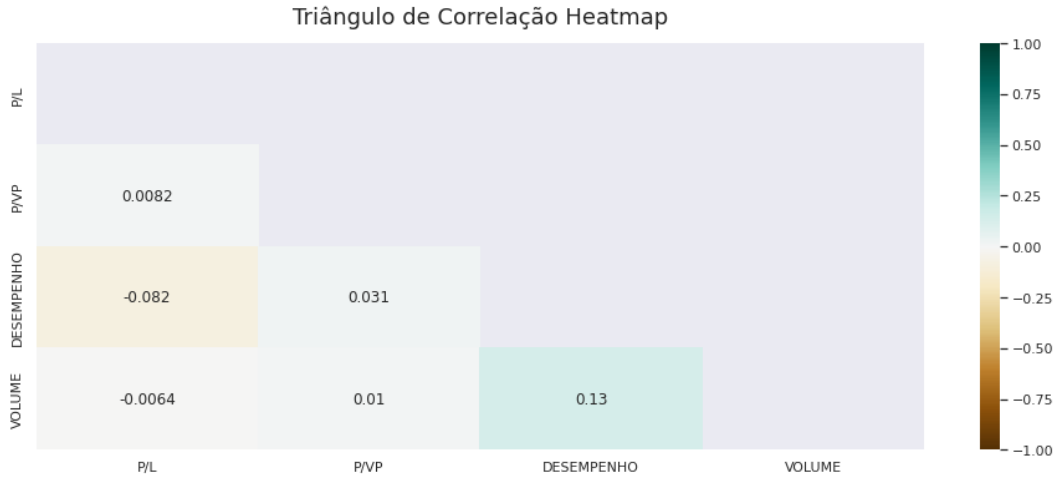

Figura 7 – Correlação dos dados - 2021.

3.3 Tratamento de Dados

Os dados utilizados são referentes a abertura e fechamento de cada ano, de 2012 a 2021, ou seja, aos valores de abertura em janeiro de 2012 e fechamento em dezembro de 2012, rendendo 9 bases de dados para execução do trabalho. Para os atributos de entrada P/L e P/VP foram utilizados os valores referentes a janeiro de cada ano, já para o outro atributo de entrada, VOL, foi utilizado o volume de negociações referentes ao período do ano anterior(janeiro a dezembro) e para a saída, Desempenho, foi calculado o percentual da diferença entre os valores de abertura (janeiro) e fechamento (dezembro) do ano anterior referente a cada ação.

Vale destacar que o índice S&P500 é dinâmico, ou seja, novas ações podem entrar ou sair do portfólio, caso uma empresa passe a cumprir ou deixar de cumprir os requisitos para compor o índice, sendo assim, iremos analisar empresas que se mantiveram compondo o índice durante todo o período testado, janeiro de 2012 à março de 2022. E com isso obtivemos um total de 302 ações para cada base de dados, mas em todas elas, sempre as mesmas empresas.

Para o tratamento do conjunto de dados foi utilizada a biblioteca Pandas, da linguagem Python, com ela pudemos organizar em colunas os valores dos atributos mencionados na seção 4.1.

Primeiramente, com os dados organizados em colunas, foram selecionados os valores referentes aos períodos de janeiro de 2012 à dezembro de 2021, após essa primeira filtragem foram retiradas linhas com valores nulos (NAN). Em seguida, foram selecionadas as ações que apareciam em todas as nove base de dados.

A Tabela 1 mostra a base de dados após o tratamento.

	NOME	P/L	P/VP	DESEMPENHO	VOLUME	ANO
0	A	12.6008	2.0698	0.142618	4601144	2013
1	AAP	13.8495	4.3877	0.017009	1130400	2013
2	AAPL	11.7256	3.8128	0.298417	527988921	2013
3	ABMD	36.2115	3.6696	-0.294488	632058	2013
4	ABT	26.8939	1.9808	0.157448	13904751	2013
...	...	...	...	...	...	...
297	WY	50.8833	4.0460	0.458071	4523024	2013
298	XRAY	20.6474	2.6035	0.110768	863800	2013
299	XYL	15.8810	2.4265	0.038314	928802	2013
300	ZBH	12.1818	2.0045	0.223568	1412458	2013
301	ZION	22.4522	0.8174	0.284514	2854523	2013

Tabela 1 – Base de dados após tratamento.

As bases de dados criadas para alimentar os modelos foram organizadas de duas formas para que se pudesse analisar as saídas por dois vieses: o primeiro quando os modelos fossem treinados com dados referentes ao período de um ano e o segundo quando os modelos fossem treinados com dados referentes ao período de três anos. Ou seja, no primeiro caso foi atribuído a variável de treino o conjunto de dados referente a 1 ano, e a variável de teste aos dados do ano seguinte, por exemplo, atribuiu-se como treino o Data Frame referente ao ano 2013, com total de 302 dados para treinamento e para teste o Data Frame referente ao ano 2014, totalizando os também 302 dados para teste e assim foi-se deslizando os anos um a um. No segundo caso foi concatenado três Data Frames para treino e usado um para teste, por exemplo, concatenou-se os Data Frames referentes aos anos de 2013, 2014 e 2015, agora totalizando 906 dados para treinamento e para teste usou-se o Data Frame do ano 2016, totalizando 302 dados para teste.

Dessa forma foi possível observar como os modelos se comportavam com diferentes quantidade de dados para treino.

3.4 Geração de modelos de regressão

O objetivo do uso dos modelos de regressão deu-se pelo fato de que eles proporcionam a predição do desempenho de ações no fechamento do ano usando como atributos de entrada P/L, P/VP do início do ano e VOL do ano anterior.

3.4.1 Modelos gerados

Para execução deste trabalho foram escolhidos modelos de regressão descritos no capítulo 2, são eles, Regressão Linear (RL), Árvore de Regressão (AR), Random Forest (RF), K-Nearest Neighbors (KNN), Bayesian Ridge (BR) e Sistemas de Inferência Fuzzy

(SIF). A escolha desses modelos deu-se pela razão de ter uma seleção variada de modelos conhecidos.

Os hiperparâmetros são parâmetros passadas ao modelo ou algoritmos visando otimizar-los, o que resultará numa melhor acurácia do modelo e devem ser definidos pelo próprio desenvolvedor da aplicação. Para esse trabalho os modelos RL, AR, RF, KNN e BR foram importados diretamente da biblioteca Scikit-learn (PEDREGOSA et al., 2011) e os parâmetros utilizados foram os *default*, uma vez foram considerados coerentes com o objetivo final, e é interessante observar que para o KNN foi usado $k = 5$. Na Figura 8 pode-se observar os valores dos hiperparâmetros dos modelos.

```
#Hiperparâmetros default dos modelos utilizados

#Regressão linear
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, copy_X=True, n_jobs=None, positive=False)

#Arvore de regressão
class sklearn.tree.DecisionTreeRegressor(*, criterion='squared_error', splitter='best', max_depth=None,
                                         min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,
                                         max_features=None, random_state=None, max_leaf_nodes=None,
                                         min_impurity_decrease=0.0, ccp_alpha=0.0)

#Random forest
class sklearn.ensemble.RandomForestRegressor(n_estimators=100, *, criterion='squared_error', max_depth=None,
                                             min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,
                                             max_features=1.0, max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True,
                                             oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False,
                                             ccp_alpha=0.0, max_samples=None)

#KNN
class sklearn.neighbors.KNeighborsRegressor(n_neighbors=5, *, weights='uniform', algorithm='auto', leaf_size=30, p=2,
                                            metric='minkowski', metric_params=None, n_jobs=None)

#Baeysian ridge
class sklearn.linear_model.BayesianRidge(*, n_iter=300, tol=0.001, alpha_1=1e-06, alpha_2=1e-06, lambda_1=1e-06, lambda_2=1e-06,
                                         alpha_init=None, lambda_init=None, compute_score=False, fit_intercept=True, normalize='deprecated',
                                         copy_X=True, verbose=False)
```

Figura 8 – Hiperparâmetros utilizados nos modelos.

O modelo baseado em regras fuzzy exige mais etapas de modelagem e por isso será explicado separadamente na seção seguinte.

3.4.2 Sistema de Inferência Fuzzy

Sistemas de Inferência Fuzzy (SIF) são sistemas compostos por regras fuzzy, que são regras no formato SE-ENTÃO que definem relações entre variáveis de entrada e de saída. Modelos de inferência como esses utilizam variáveis linguísticas, que são variáveis cujos valores são termos linguísticos ao invés de números, sendo que cada termo linguístico está associado a um conjunto fuzzy.

Esse modelo foi incluído nas análises com o objetivo de explorar a representação do conhecimento na forma de termos linguísticos e regras, visando compreender o significado do modelo.

Neste trabalho a implementação do modelo Fuzzy foi feita na linguagem Python e para a função de pertinência foi escolhida a representação pela forma triangular, em que para

os atributos de entrada do modelo, ou seja, P/L, P/VP e VOL seus respectivos domínios foram granularizados em três categorias representadas pelos termos Baixo, Médio e Alto. Já para o atributo que deseja-se prever, o desempenho, o domínio foi granularizado em 5 categorias representadas pelos termos Muito Baixo, Baixo, Médio, Alto e Muito Alto.

Os parâmetros das funções triangulares foram definidos manualmente são diferentes para cada atributo. Esses parâmetros foram ajustados manualmente com base nos gráficos de dispersão dos atributos, em que foi observada uma concentração dos dados e assim foram feitos experimentos preliminares visando encontrar o formato e distribuição mais significativos para cada atributo. Nas Tabelas 2 e 3 pode-se observar esses valores para as entradas (P/L, P/VP, VOL) e a saída (desempenho).

	Baixo	**Médio**	**Alto**
P/L	[0, 0, 20]	[15, 25, 35]	[20, 40, 40]
P/VP	[2, 2, 4]	[2, 4, 10]	[4, 10, 10]
VOL(10^6)	[1, 1, 5]	[1, 5, 12]	[5, 12, 12]

Tabela 2 – Parâmetros das funções triangulares.

	Muito Baixo	**Baixo**	**Médio**	**Alto**	**Muito Alto**
Dsp	[-0.3, -0.3, 0.03]	[-0.1, 0.075, 0.14]	[0.05, 0.125, 0.2]	[0.125, 0.2, 0.4]	[0.25, 0.5, 0.5]

Tabela 3 – Parâmetros das funções triangulares do atributo de saída (desempenho).

Após a definição dos intervalos foram montados os triângulos. Um ponto importante a se destacar é que para os triângulos das extremidades, ou seja, Baixo e Alto para os atributos de entrada e Muito Baixo e Muito Alto para o de saída foi atribuído o grau de pertinência 1 quando os valores dos elementos ficassem fora dos valores dos intervalos. Esse tratamento fez-se necessário para que os outliers fossem contemplados pelo modelo. As Figuras 9, 10, 11, 12 ilustram os triângulos.

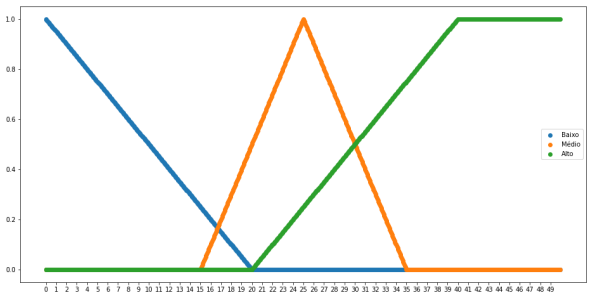

Figura 9 – Função triangular - P/L.

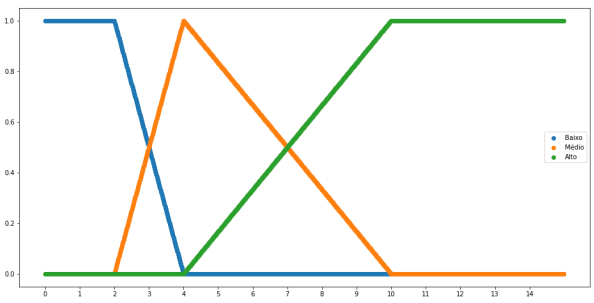

Figura 10 – Função triangular - P/VP.

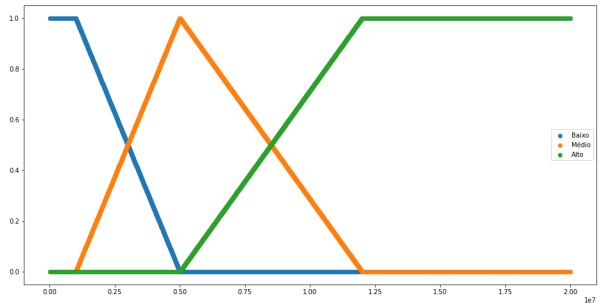

Figura 11 – Função triangular - VOL.

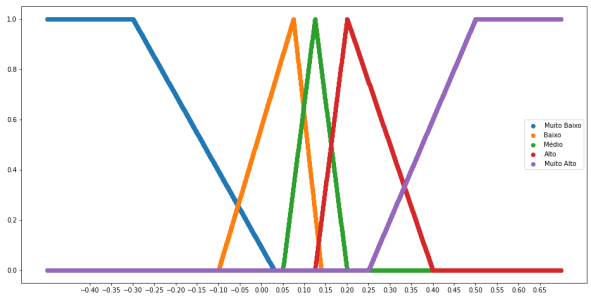

Figura 12 – Função triangular - Desempenho.

Com a definição dos conjuntos fuzzy foi realizado o processo de granularização dos dados, em que as entradas com valores numéricos foram transformadas em atributos categóricos. Para tal processo, foram calculados os graus de pertinência dos valores de cada atributo e a substituição desse valor numérico pelo termo linguístico que representa o conjunto com maior grau de pertinência. Na Figura 13 pode-se observar o novo conjunto de dados após a granularização.

	P/L	P/VP	VOLUME	DESEMPENHO
0	BAIXO	BAIXO	MEDIO	MEDIO
1	BAIXO	MEDIO	BAIXO	BAIXO
2	BAIXO	MEDIO	ALTO	ALTO
3	ALTO	MEDIO	BAIXO	MUITO BAIXO
4	MEDIO	BAIXO	ALTO	MEDIO
...	...	...	...	...
297	ALTO	MEDIO	MEDIO	MUITO ALTO
298	MEDIO	BAIXO	BAIXO	MEDIO
299	BAIXO	BAIXO	BAIXO	BAIXO
300	BAIXO	BAIXO	BAIXO	ALTO
301	MEDIO	BAIXO	BAIXO	ALTO

Figura 13 – Entradas em formato categórico.

Fonte: autoria própria.

O novo conjunto de dados, agora com os atributos em formato categórico, foi utilizado como entrada para o cálculo da base de regras por meio do algoritmo de árvore de decisão C4.5. utilizando a implementação de árvore de decisão J48 com auxilio da ferramenta Weka (https://www.cs.waikato.ac.nz/ ml/weka/), e com ela foi extraída toda base de regras. A ferramenta Weka foi escolhida uma vez que a implementação do modelo de árvore de decisão do scikit-learn não suporta atributos categóricos, apenas numéricos. É

importante salientar que a árvore de decisão é um modelo para classificação e foi usada apenas para extração das regras fuzzy. Como o processo de inferência utilizado é um processo de inferência fuzzy, a saída do modelo, após a defuzificação, é um valor numérico, portanto adequado a problemas de regressão.

Na Figura 14 pode-se observar a árvore de decisão gerada sem poda, em que cada folha da árvore dá origem a uma regra, formada pelos atributos que aparecem no caminho desde a raiz até a folha.

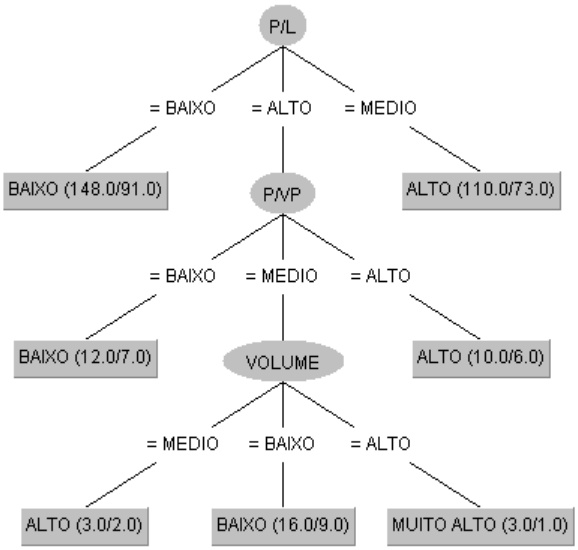

Figura 14 – Árvore de Decisão.

Com a base de regras fuzzy extraídas da árvore de decisão, foi aplicado o método descrito no capítulo 2 (seção 2.5.3). No método de Mamdani, os valores numéricos dos atributos de entrada são apresentados ao modelo, que calcula o grau de pertinência de cada um deles no conjunto fuzzy correspondente de cada regra e o grau de disparo das regras calculado pela conjunção pelo mínimo. As saídas das regras disparadas são agregadas pelo operador de união máximo e, na sequência, pode-se realizar o processo de defuzificação, ou seja, converter a saída final agregada para um valor numérico. Para isso foi utilizado o método Centro dos Máximos Ponderado e assim foram obtidos os valores preditos para o atributo desempenho.

3.5 Construção das Carteiras

Seguindo o objetivo proposto pelo trabalho, foram construídas as carteiras de investimento para os dois cenários mencionados anteriormente: Cenário 1: dados de um ano para geração do modelo, dados do ano seguinte para predição do desempenho; Cenário

2: dados de três anos consecutivos para geração do modelo, dados do ano seguinte para predição do desempenho. 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.

Para cada um dos modelos foram organizadas todas as ações de maneira decrescente de acordo com o desempenho predito e então selecionadas as 15 primeiras ações, ou seja, as que tiveram melhores desempenhos. Posteriormente foi calculada a média aritmética do das ações do conjunto selecionado para obter o valor do desempenho anual da carteira para cada modelo.

Além disso, para fins de comparação, foi calculado também a média aritmética dos valores preditos do conjunto inteiro de ações para os dois cenários, primeiro com as 302 e segundo com as 906 ações. Para esse caso, deu-se o nome de 'S&P500 Modificado'.

Capítulo 4

Experimentos e análise dos resultados

Nesse capítulo serão descritos os experimentos realizados e apresentados os resultados das predições dos modelos, com avaliação do erro de predição além das análises de eficácia da estratégia proposta para a construção das carteiras de investimentos. Sendo assim será possível avaliar qual modelo obteve melhor performance, como também se algum cenário de treinamento e teste se sobrepôs ao outro.

O Python é uma linguagem de programação amplamente difundida e que possui várias bibliotecas auxiliares e na área de AM não é diferente. Nesse trabalho será utilizada a biblioteca Scikit-learn, que apresenta uma gama variada de modelos e uma praticidade em sua implementação. Também foi utilizada a ferramenta Weka para para geração da árvore de decisão para o modelo fuzzy.

4.1 Avaliação dos modelos

A avaliação dos modelos é uma parte fundamental para identificar se os algoritmos implementados estão comportando-se de maneira desejada. Para este trabalho foram utilizados o EQM e a Validação Cruzada (VC).

Na VC os dados iniciais são particionados aleatoriamente em k subconjuntos de tamanho aproximadamente igual em que o treinamento e o teste são realizados k vezes. A cada iteração do modelo um subconjunto diferente é reservado para teste, enquanto os outros são utilizados para o treinamento. Ao contrário dos métodos de validação por subamostragem aleatória, aqui cada amostra é usada o mesmo número de vezes para treinamento

e uma vez para teste. Dessa forma, pode-se estimar com mais exatidão quão preciso é o modelo na prática (HAN; KAMBER; PEI, 2012).

Para este trabalho os modelos foram avaliados pelo EQM calculado para os dados de teste e também para os dados de treinamento, visando verificar se os modelos sofreram sobre ajuste. Na avaliação com os dados de treinamento foi utilizada a validação cruzada com k = 5.

Para melhor análise dos resultados, os métodos de regressão foram organizados de maneira crescente do valor de EQM a cada ano, e para cada colocação foram dados valores numéricos de 1 a 6, com 1 para o menor valor e 6 para o maior valor. Após a organização em ordem crescente e atribuição de valores numéricos foram feitas as médias aritméticas das colocações dos modelos ao longo dos anos, e assim conseguiu-se obter valores que pudessem expressar quais modelos apresentarem menores valores para EQM.

4.1.1 Cenário 1

A Tabela 4 mostra os valores dos EQM dos modelos RL, RF, KNN, BR e AR dos anos de 2014 a 2021. Os dados foram organizados de forma que primeiro número do vetor é o EQM da predição com os dados de teste (ano seguinte) e o segundo número é o EQM da predição com dados de treinamento usando validação cruzada

Analisando-se a figura observa-se que os modelos não sofreram sobre ajuste, uma vez que os valores dos EQM dos dados de teste e de treinamento foram próximos.

Vale destacar que para o cenário 1 não foi implementado o modelo fuzzy, sendo assim, não será feita nenhuma análise sobre ele nessa seção.

	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)

Tabela 4 – Erros quadráticos médios das predições com dados de treinamento e teste - cenário 1

Na Tabela 5 pode-se observar o ranking dos EQM dos modelos 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.

	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**

Tabela 5 – Médias dos erros quadráticos médios - cenário 1

4.1.2 Cenário 2

Assim como no cenário 1 a Figura 6 mostra os valores dos EQM dos modelos RL, RF, KNN, BR, AR e SIF, mas dessa vez dos anos de 2016 a 2021. Os dados foram organizados de forma que primeiro número do vetor é o EQM da predição com os dados de teste e o segundo número é o EQM da predição com dados de treinamento usando validação cruzada.

Analisando-se a Tabela 6 observa-se que para esse cenário os modelos também não sofreram sobre ajuste, uma vez que os valores dos EQM de treinamento e teste foram próximos.

	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)

Tabela 6 – Erros quadráticos médios das predições com dados de teste e de treinamento - cenário 2

Na tabela 7 pode-se observar o ranking dos EQM 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 cenários.

	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**

Tabela 7 – Médias dos erros quadráticos médios - cenário 2

4.2 Apresentação das Carteiras

Para avaliar a estratégia de alocação de carteiras, foram geradas carteiras com as 15 ações que obtiveram o melhor desempenho predito por cada uma dos modelos. O desempenho predito geral de cada carteira pode então ser comparado com o desempenho real. Os desempenhos obtidos com as carteiras criadas com o uso dos modelos de regressão foram também comparados com um método de construção de carteira manual proposto por Graham (2016) e com o índice S&P500 Ajustado, descrito no capítulo 2.

O primeiro passo foi a construção da carteira manual, para isso foi aplicado o método de Graham (2016) e seguiu-se a seguinte regra: multiplicou-se os indicadores P/L e P/VP e então ordenou-se as ações de maneira crescente de acordo com o resultado da multiplicação, a partir disso foram selecionadas as 15 primeiras ações, ou seja, as com menor valor na relação $\mathrm{P/L} * \mathrm{P/VP}$, posteriormente foi feita a média aritmética do desempenho dessas 15 ações, e assim obteve-se o desempenho anual da carteira manual.

As Tabelas 8 e 9 apresentam os resultados das carteiras para os cenários 1 e 2 respectivamente, sendo o desempenho real os valores verdadeiros de quanto o conjunto das 15 ações selecionadas representou naquele ano, e o desempenho predito os valores previstos de quanto o conjunto das 15 ações selecionadas representou também naquele ano. É importante destacar que para cada ano e modelo as carteiras são distintas, ou seja, são compostas por ações diferentes. Os números estão representados percentualmente, pois eles representam o percentual de valorização ou desvalorização que a carteira obtive no ano. Os desempenhos reais estão estampados com a cor azul, enquanto os desempenhos preditos com a vermelha.

Já as Figuras 15, 16, 17, 18 mostram a representação gráfica dos desempenhos preditos e reais para os cenários 1 e 2.

	2014	2015	2016	2017	2018	2019	2020	2021	Média
Cart. Manual	10.0%	-1.8%	29.7%	18.6%	-17.1%	30.6%	-7.2%	28.3%	11.3%
S&P500 Ajustado	16.4%	2.1%	17.3%	20.8%	-6.7%	32.5%	12.1%	27.5%	15.2%
Reg. Linear	35.3% 44.0%	28.1% 19.7%	7.4% 34.7%	17.7% 21.8%	-10.0% 23.5%	39.3% 6.1%	31.9% 45.0%	28.5% 140%	22.3% 41.9%
Rand. Forest	20.6% 71.3%	33.6% 40.6%	27.3% 42.0%	25.1% 55.6%	4.3% 48.6%	44.5% 12.1%	36.9% 57.6%	35.0% 64.2%	28.4% 49.0%
KNN	12.4% 63.9%	1.2% 42.7%	19.6% 38.6%	33.8% 50.6%	-12.5% 46.3%	31.8% 14.5%	26.9% 55.7%	36.1% 39.9%	18.7% 44.0%
Bacys. Ridge	35.3% 43.5%	25.3% 18.9%	7.4% 31.8%	23.1% 21.6%	-10.0% 23.5%	39.4% 4.8%	18.3% 41.2%	28.5% 131%	20.9% 39.6%
Arv. Regres.	16.9% 108%	12.0% 76.6%	27.0% 58.5%	19.6% 83.1%	-7.2% 85.5%	37.5% 36.3%	21.9% 79.6%	32.8% 94.0%	20.1% 77.8%

Tabela 8 – Desempenhos reais e preditos pelas carteiras geradas por cada modelo.

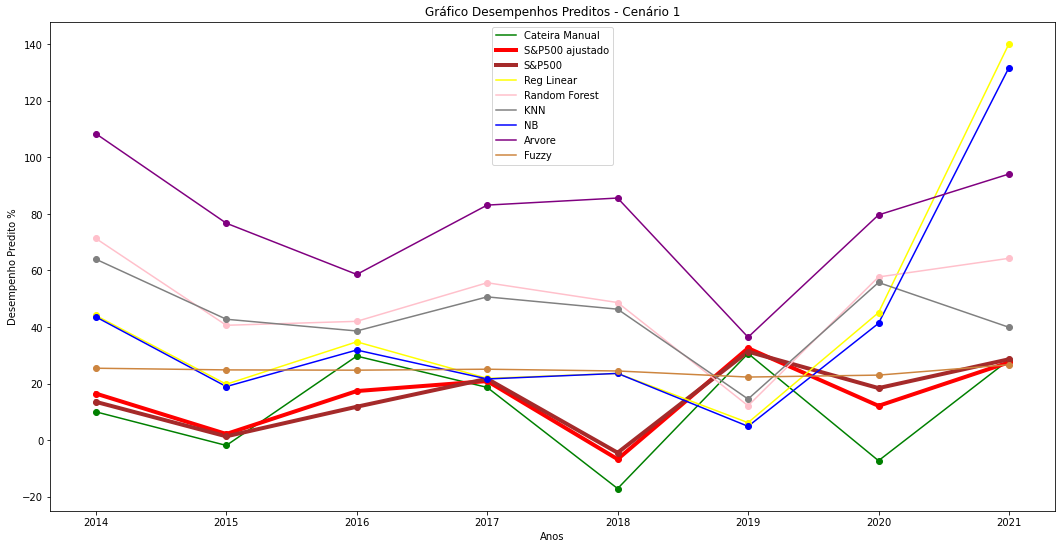

Figura 15 – Desempenhos preditos pelas carteiras geradas por cada modelo.

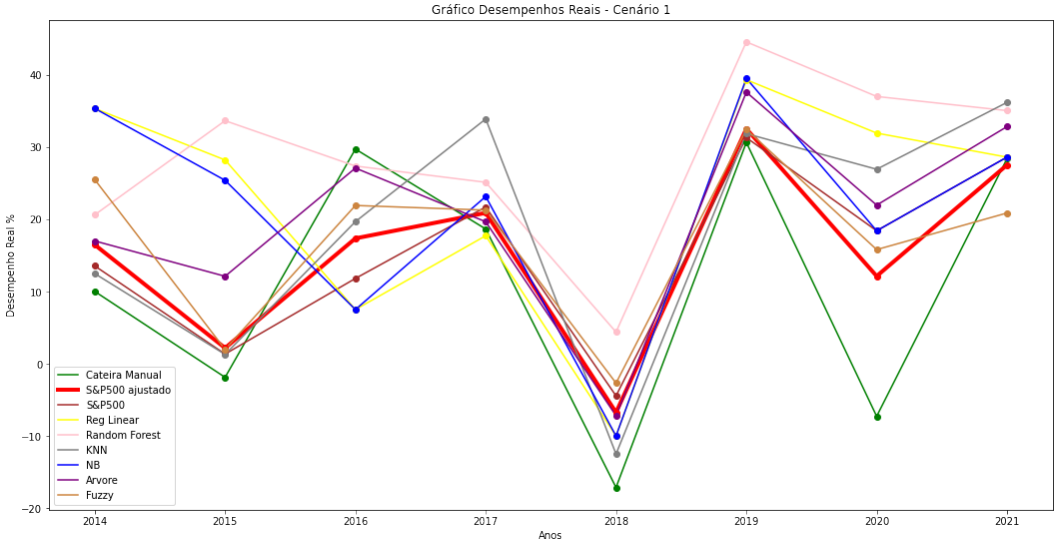

Figura 16 – Desempenhos reais pelas carteiras geradas por cada modelo.

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 acima da média do índice, com a árvore de regressão obtendo o melhor desempenho, 77.8%, enquanto BR apresentou menor desempenho, 20.9%, sendo que apenas a Carteira Manual obteve um resultado abaixo do índice S&P500 Ajustado, 11.3%. Quando olha-se para os desempenhos reais, ou seja, o valor médio do desempenho real correspondente as 15 ações selecionadas pela predição constata-se também que todos os modelos 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.

Outro ponto interessante é 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 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 acima da média e constantes.

	2016	2017	2018	2019	2020	2021	Média
Cart. Manual	37.3%	17.0%	20.1%	22.0%	22.8%	5.2%	20.7%
S&P500 Ajustado	17.3%	20.8%	-6.7%	32.5%	12.1%	27.5%	17.2%
Reg Linear	27.5%	26.2%	3.8%	44.0%	41.2%	20.6%	27.2%
(Real / Pred)	24.5%	32.9%	26.3%	23.9%	28.4%	154.4%	48.4%
Rand Forest	10.0%	40.4%	-0.19%	41.7%	51.8%	19.0%	27.3%
(Real / Pred)	54.1%	56.0%	43.2%	45.9%	60.0%	63.0%	53.7%
KNN	20.9%	16.5%	-9.1%	43.4%	20.0%	26.3%	19.7%
(Real / Pred)	67.1%	46.6%	35.9%	48.3%	33.4%	43.0%	45.7%
Bacys. Ridge	28.2%	26.5%	3.8%	44.0%	40.1%	20.6%	27.2%
(Real / Pred)	24.3%	29.3%	25.6%	23.4%	27.4%	151.1%	46.8%
Arv. Regressão	15.4%	39.5%	-5.6%	38.6%	41.3%	30.7%	26.6%
(Real / Pred)	103.4%	97.9%	88.3%	66.3%	77.3%	80.9%	85.7%
Fuzzy	2.7%	30.8%	-6.3%	37.5%	29.7%	37.5%	22.0%
(Real / Pred)	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50%

Tabela 9 – Desempenhos reais e preditos pelas carteiras geradas por cada modelo.

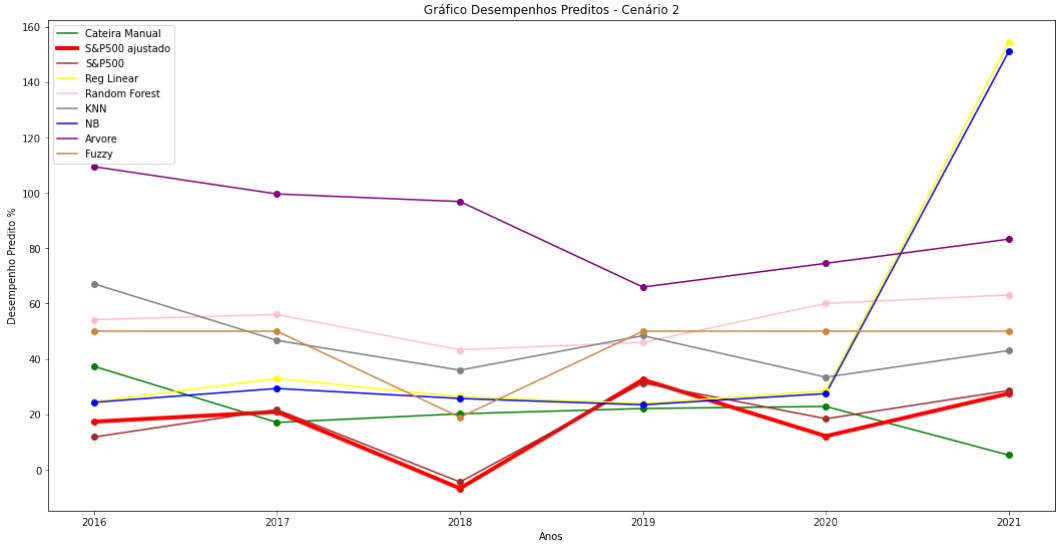

Figura 17 – Desempenhos preditos pelas carteiras geradas por cada modelo.

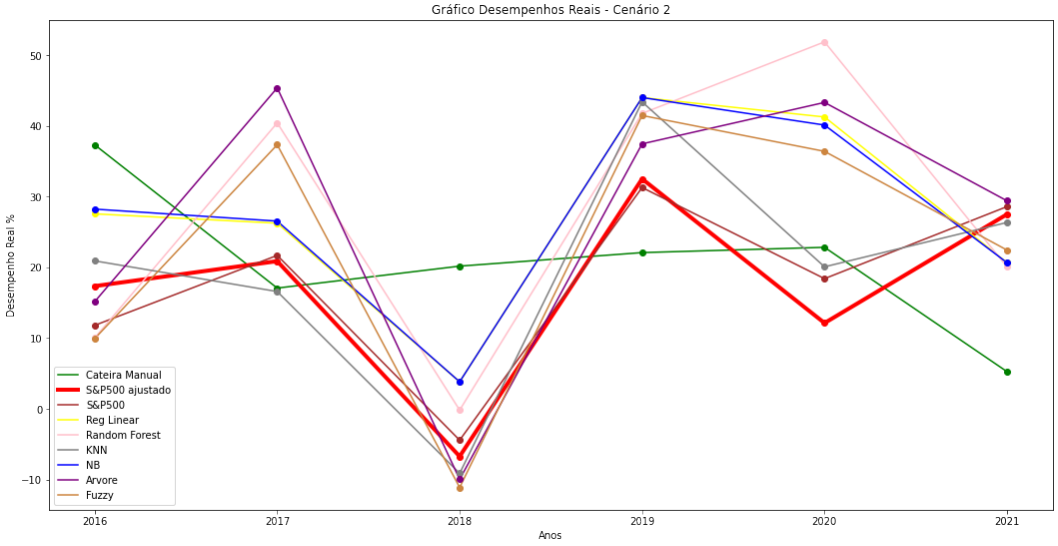

Figura 18 – Desempenhos reais pelas carteiras geradas por cada modelo.

Para o cenário 2 é importante frisar que há o acréscimo do SIF na construção da carteira e 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 AM obtiveram desempenhos acima da média do índice, com a AR obtendo o melhor desempenho, 85.7%, enquanto KNN apresentou menor desempenho, 45.7%. Quando olha-se para os desempenhos reais, ou seja, o valor médio do desempenho real correspondente as 15 ações selecionadas pela predição constata-se também que todos os modelos 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 quando comparado ao 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.

Também 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 ambos os modelos, com isso constata-se que os modelos apresentaram estabilidade, como também desempenho acima da média e constantes.

4.3 Interpretabilidade das regras fuzzy

Como já descrito no capítulo anterior, para este trabalho as regras fuzzy utilizadas para a construção do modelo foram extraídas por meio da árvore de decisão J48, para posteriormente serem usadas no processo de inferência.

Na figura 19 pode-se observar a árvores de decisão gerada com poda, em que cada folha da árvore dá origem a uma regra, formada pelos atributos que aparecem no caminho desde a raiz até a folha. A árvore em questão foi gerada pelo conjunto dados referentes aos anos 2013 à 2015.

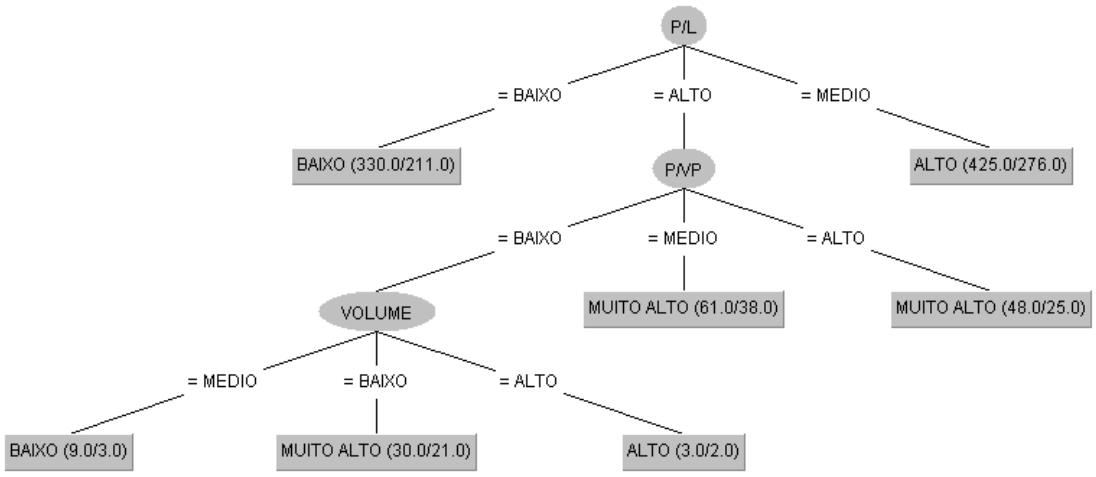

Figura 19 – Árvore de decisão do conjunto de treino 2013-2014-2015.

Por meio da árvore foi extraído o conjunto de regras listado a seguir, podendo-se observar que no geral para P/L alto ou médio o desempenho é "alto"ou "muito alto", como é mostrado nos regras 5, 6 e 7. Já o valor do P/VP não teve influência direta na saída, uma vez que para entrada "baixo"o desempenho resultante oscilou entre "baixo", "alto"e "muito alto". Em relação ao volume também não é verificada um relação com o desempenho resultante.

1. Se P/L BAIXO então DESEMPENHO BAIXO
2. Se P/L ALTO E P/VP BAIXO e VOLUME MÉDIO então DESEMPENHO BAIXO
3. Se P/L ALTO e P/VP BAIXO e VOLUME BAIXO então DESEMPENHO MUITO ALTO
4. Se P/L ALTO e P/VP BAIXO e VOLUME ALTO então DESEMPENHO ALTO

5. Se P/L ALTO e P/VP MÉDIO então DESEMPENHO MUITO ALTO
6. Se P/L ALTO e P/VP ALTO então DESEMPENHO MUITO ALTO
7. Se P/L MÉDIO então DESEMPENHO ALTO

Capítulo 5

Conclusão e trabalhos futuros

Conclusão

Este trabalho de conclusão de curso teve como objetivo a utilização de modelos de AM 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 ao próprio índice de mercado.

Foram utilizados sete modelos de AM para predizer os valores de um conjunto de ações pertencentes ao índice S&P500, são eles os modelos: Regressão Linear , Árvore de Regressão, Random Forest, K-Nearest Neighbors , Bayesian Ridge e Sistema de inferência fuzzy além da construção das carteiras pelo método de Graham (2016). A avaliação dos modelos foi feita através do cálculo do erro quadrático médio.

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.

Os objetivos do trabalho foram cumpridos, uma vez que conseguiu-se propor modelos eficazes e simples para a construção de carteiras de investimentos e que fossem capazes de ter desempenhos superiores as médias de mercado e ao método manual.

Trabalhos Futuros

A forma na qual o conjunto de dados foi selecionado, com um período de aproximadamente 10 anos, pode ter influenciado nos resultados finais de desempenho dos modelos, uma vez que nesse período, 2012 a 2020, não houve nenhuma crise econômica significativa

que pudesse alterar bruscamente os atributos de entrada, de forma que alteraria o comportamento dos modelos, sendo assim uma abordagem diferente daria-se por alimentar os modelos com um conjunto de dados referentes a um período maior que dez anos.

Além disso, para treinamento dos modelos foram utilizados conjuntos de dados com 3 indicadores fundamentalistas como entrada, P/L, P/VP e VOL, uma abordagem com um maior número de indicadores, como valores de Dividendos anuais, Ebitda, ROE, dentre outros, pode trazer mais informações sobre as empresas e consequentemente uma maior assertividade para os modelos.

Outra abordagem possível é a mudança de granularização dos conjuntos Fuzzy, alterando os intervalos dos triângulos como também aumentar o número de intervalos. Além disso, utilizar diferentes funções de representação dos conjuntos, como as formas Trapezoidal e Gaussiana.

Referências

BASU, S. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. **The journal of Finance**, Wiley Online Library, v. 32, n. 3, p. 663–682, 1977.

_____. The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. **Journal of financial economics**, Elsevier, v. 12, n. 1, p. 129–156, 1983.

CASTRO, S. R. d. C. **Alocação de carteiras de ações através da utilização de modelos de lógica Fuzzy**. Tese (Doutorado), 2009.

FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **The journal of Finance**, JSTOR, v. 25, n. 2, p. 383–417, 1970.

FRANCHI, B. O. Análise comparativa das metodologias de markowitz, kelly e aprendizado por reforço em carteiras de investimentos-uma abordagem computacional. Universidade Federal de São Paulo, 2021.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. [S.l.]: "O'Reilly Media, Inc.", 2019.

GRAHAM, B. **O investidor inteligente**. [S.l.]: HarperCollins Brasil, 2016.

HAN, J.; KAMBER, M.; PEI, J. 8 - classification: Basic concepts. In: HAN, J.; KAMBER, M.; PEI, J. (Ed.). **Data Mining (Third Edition)**. Third edition. Boston: Morgan Kaufmann, 2012, (The Morgan Kaufmann Series in Data Management Systems). p. 327–391. ISBN 978-0-12-381479-1. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780123814791000083>.

INDICES, S. D. J. S&p dow jones indices. **Retrieved April**, v. 12, p. 2016, 2016. Disponível em: <https://www.spglobal.com/spdji/en/documents/index-news-and-announcements/20161110-sp-500-minimum-volatility-consultation-results.pdf>.

JANG, J.-S.; SUN, C.-T. Predicting chaotic time series with fuzzy if-then rules. In: IEEE. **[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems**. [S.l.], 1993. p. 1079–1084.

JUNIOR, J. G. A. S. **Um Estudo Sobre Aprendizado de Máquina Aplicado à Modelagem de Retornos de Ações**. Dissertação (Mestrado) — Brasil, 2015.

KLIR, G.; YUAN, B. **Fuzzy sets and fuzzy logic**. [S.l.]: Prentice hall New Jersey, 1995. v. 4.

LUGER, G. F. **Inteligência Artificial-: Estruturas e estratégias para a solução de problemas complexos**. [S.l.]: Bookman, 2004.

MACKAY, D. J. Bayesian interpolation. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 4, n. 3, p. 415–447, 1992.

MAMDANI, E. H.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. **International journal of man-machine studies**, Elsevier, v. 7, n. 1, p. 1–13, 1975.

MELO, C. **Regressão Linear: Conceitos e Implementação com Python**. 2022. Urlhttps://sigmoidal.ai/como-implementar-regressao-linear-com-python/. Acesso em: 20 jun. 2022.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal**, v. 24, n. 3, p. 69–71, 2012.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PIMENTA, A. H. d. M. Geração genética de classificador fuzzy intervalar do tipo-2. Universidade Federal de São Carlos, 2009.

ROSS, S. A. Neoclassical finance. In: **Neoclassical Finance**. [S.l.]: Princeton University Press, 2009.

RUSSELL, S. J. **Artificial intelligence a modern approach**. [S.l.]: Pearson Education, Inc., 2010.

TIPPING, M. E. Sparse bayesian learning and the relevance vector machine. **Journal of machine learning research**, v. 1, n. Jun, p. 211–244, 2001.

TURING, A. M.; HAUGELAND, J. Computing machinery and intelligence. **The Turing Test: Verbal Behavior as the Hallmark of Intelligence**, p. 29–56, 1950.

YAGUINUMA, C. A. Processamento de conhecimento impreciso combinando raciocínio de ontologias fuzzy e sistemas de inferência fuzzy. Universidade Federal de São Carlos, 2013.

ZADEH, L. Fuzzy sets, 1965 fuzzy sets. **Informationand control**, p. 338–338, 1965.