Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Aplicação de árvores de decisão na seleção de portfólios de ações

Vitor Eduardo Galeão Borba de Borba

2 de junho de 2021

Vitor Eduardo Galeão Borba de Borba

Aplicação de árvores de decisão na seleção de portfólios de ações

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra. Taiane Schaedler Prass
Coorientador: Prof. Dr. Marcio Valk

Porto Alegre
19 de maio de 2021

Vitor Eduardo Galeão Borba de Borba

Aplicação de árvores de decisão na seleção de portfólios de ações

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pelo(a) Orientador(a) e pela Banca Examinadora.

Orientadora:_____
Profa. Dra. Taiane Schaedler Prass, UFRGS
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Hudson Torrent, UFRGS
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Prof. Dr. Marcio Valk, UFRGS
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Porto Alegre
19 de maio de 2021

"Se ninguém pudesse transacionar, se cada indivíduo fosse forçado a ser totalmente autossuficiente, a maioria de nós obviamente morreria de fome, e o restante mal conseguiria se manter vivo. A troca é a força vital não só da economia, mas da própria civilização." (Murray N. Rothbard)

Agradecimentos

Agradeço a Deus, o autor da vida, em quem mantive firme minha fé e que me sustentou ao longo do curso.

Aos meus pais Jairo e Vera, por ensinarem tudo que eu sou e por serem sempre um exemplo para mim, tanto de família quanto de ensino.

À minha amada esposa, Raquel, pelo seu sorriso a cada manhã e por me incentivar e estar sempre comigo para o que der e vier. *Ich liebe Dich...*

Aos meus colegas de curso e universidade, que estiveram mais próximos em diversos momentos ao longo desses anos, como Aline Foerster, Aline Gularte, Bernardo Altenbernd, Cristiano Sulzbach, Douglas Lopes, Eduardo de Oliveira, Gabriela Rech, Lincon Camargo, Luciano Reis, Luís de Almeida, Mariana Garcia, Monika Sohne, Tobias Gomes. Todos vocês, em algum período da faculdade, tiveram uma participação muito importante para mim.

À professora Taiane, minha orientadora. Obrigado pela confiança, imenso apoio, carinho recebido em forma de atenção e ajuda, independentemente da hora em que pedia.

Ao professor Marcio por me coorientar e me apresentar para a minha orientadora Taiane e, juntamente com o professor Hudson, serem banca deste trabalho.

Ao professor Fernando Pulgatti pelos ensinamentos passados na bolsa no NAE-UFRGS.

Às professoras Debora Feijó, Lurdes Busin, Karina Azzolin e a toda equipe de Enfermagem em Terapia Intensiva do Hospital de Clínicas de Porto Alegre - HCPA, pela confiança e pelos ensinamentos transmitidos no estágio.

Às empresas Viação Teresópolis Cavalhada e Banco Agibank por abrirem as portas através do estágio e possibilitarem aplicar os conhecimentos adquiridos na estatística.

E, por fim, gostaria de agradecer a todos meus familiares e amigos que, de alguma forma, contribuíram para a chegada deste momento.

Resumo

A teoria moderna de portfólio baseia-se na ideia de diversificação dos ativos para otimizar as carteiras de investimento. Entretanto, as metodologias mais utilizadas na literatura partem do pressuposto de normalidade dos dados, ou, ainda, assumem uma relação linear entre o retorno dos ativos e do mercado. O método de Markowitz, além da hipótese de normalidade, depende ainda da estimação da matriz de covariâncias e a inversão da mesma. A aplicação desse método pode ser uma tarefa difícil, quando a quantidade de ativos envolvidos é muito grande. O presente trabalho tem por objetivo explorar e aplicar árvores de decisão, uma técnica de *machine learning*, para a seleção de um portfólio de ações no mercado brasileiro. Tal metodologia surge como uma alternativa para flexibilizar a hipótese de normalidade, a estrutura de correlação dos dados e, ainda, para manipular grandes volumes de dados. A metodologia proposta é apresentada por meio de uma aplicação às séries históricas dos preços das ações listadas na bolsa brasileira (B3). Com intuito de avaliar o desempenho da carteira de investimento oriunda das árvores de decisão, utilizamos o retorno de indicadores como Selic e Ibovespa e das carteiras construídas por dois métodos clássicos: o método de Markowitz e o da carteira ingênua (*naive*).

Palavras-Chave: Portfólio, Ações, Árvore de Decisão, Aprendizado de Máquina, Markowitz, Ibovespa.

Abstract

Modern portfolio theory is based on the idea of assets diversification, in order to optimize investments. However, the most commonly used methodologies are based on the hypothesis of normality or on the assumption of a linear relationship between the market's and the assets' returns. The Markowitz method, in addition to the normality hypothesis, also depends on the estimation of the covariance matrix and its inversion. Application of this method can be a difficult task when the amount of assets involved is very large. In this work our goal is to explore and apply decision trees, a machine learning technique, for portfolio selection in the Brazilian market. Such methodology emerges as an alternative to relax the hypothesis of normality, the correlation structure of the data and also, to manipulate large volumes of data. The proposed methodology is presented through an application to the historical series of stock prices listed in the brazilian stock market (B3). In order to evaluate the performance of the portfolio derived from decision trees, we used the return of indicators such as Selic and Ibovespa and the portfolios built by two classic methods: the Markowitz method and the naive portfolio.

Keywords: Portfolio, Stocks, Decision tree, Machine Learning, Markowitz, Ibovespa.

Sumário

1 Introdução

A Bolsa de Valores é um dos espaços mais democráticos do mundo, pois nela é possível negociar ativos apenas tendo uma simples conta em uma corretora de valores, onde praticamente qualquer cidadão consegue abrir. Também é possível ser sócio de empresas investindo nelas para o longo prazo, sem mencionar que há ainda a possibilidade de comprar sem ter dinheiro e vender sem ter mercadoria nos mercados futuros, em que menos de 2% das operações são liquidadas pela entrega efetiva do bem transacionado (Fortuna, 2007, página 635). Em síntese, investir na Bolsa de Valores é fácil, porém rentabilizar esses investimentos, gerar valor e multiplicar o capital não seguem essa mesma facilidade.

Existem inúmeras técnicas e estratégias de se operar na Bolsa de Valores, dentre elas, está a de selecionar um portfólio de ações, no qual, em vez de ficar restrito apenas à compra de um ativo, dilui-se o risco diversificando a escolha dos ativos, montando uma carteira de ações onde tenham-se N ativos presentes nela. Um dos grandes desafios na hora de investir passa então a ser a escolha dos ativos que farão parte da carteira de investimentos. Outra questão importante, e tão desafiadora quanto, é determinar qual o peso que cada um desses ativos terá. Sendo assim, auxiliar os investidores na escolha dos pesos que tornam essa diversificação mais eficaz é de extrema relevância em um cenário em que queremos otimizar um portfólio de ações.

A teoria moderna de portfólio, proposta inicialmente por Markowitz (1952), usa exatamente a ideia de diversificação dos ativos para otimizar as carteiras de investimento. Embora seja amplamente utilizada, tal metodologia parte do pressuposto de normalidade dos dados. Além disso, envolve a estimação da matriz de covariância e a inversão da mesma, o que pode ser uma tarefa difícil, quando a quantidade de ativos envolvidos é muito grande.

Nesse contexto, pretende-se utilizar árvores de decisão, uma técnica de *machine learning* que vem ganhando crescente atenção na literatura, com o objetivo de selecionar ativos para compor um portfólio. A metodologia aqui proposta surge como

uma alternativa para flexibilizar a hipótese de normalidade e manipular grandes volumes de dados. A rentabilidade da carteira construída a partir da metodologia sugerida será comparada com a de carteiras obtidas com técnicas e indicadores conhecidos, como o método de Markowitz, a carteira *Naive*, a taxa Selic e o índice Ibovespa.

Neste trabalho, levaremos em consideração apenas os ativos listados na B3, conhecida anteriormente pelo nome de BM&F Bovespa, na qual estão listadas para negociação as ações das principais empresas brasileiras. Consideraremos os preços ajustados e os retornos em diferentes horizontes de tempo a partir da segunda década do século XXI, visto que o acesso às informações e aos dados após esse período começa a ficar mais trivial de serem buscados. O presente trabalho contará com o desenvolvimento do processo de seleção de carteiras de investimentos através da utilização do *software*[1] R (R Core Team, 2020) pois, além de ser um *software* livre, também possui uma ampla flexibilidade para buscar, manipular e apresentar os dados provenientes da Bolsa de Valores.

Este trabalho é organizado da seguinte forma: no Capítulo 2, apresentamos os principais conceitos relacionados a portfólios, bem como uma breve descrição das técnicas empregadas neste trabalho. No Capítulo 3, apresentamos o estudo empírico realizado e os resultados obtidos. O Capítulo 4 é dedicado às conclusões e, para finalizar, são apresentadas as referências bibliográficas.

[1] Os códigos implementados pelo autor deste trabalho estão disponíveis mediante solicitação.

2 Referencial Teórico

Como é evidenciado na literatura, é possível deparar-se com diversos estudos e constante desenvolvimento de pesquisas com o propósito de encontrar uma maneira quantitativa de mensurar e organizar um portfólio de ações, de modo a obter uma relação ótima entre risco e retorno. Veja, por exemplo, Prass (2008); Prass e Lopes (2012); Santos (2012) e referências ali contidas.

O início da teoria moderna de portfólio foi proposto por Markowitz. O método que hoje recebe seu nome tem como ideia definir qual o peso[1] que cada ativo terá na carteira de investimento, após os ativos já terem sido escolhidos. Em conjunto com o método de Markowitz, utiliza-se frequentemente o modelo CAPM (*Capital Asset Pricing Model*), que é uma ferramenta empregada para analisar o comportamento dos ativos e auxiliar na tomada de decisão de quais ativos incluir no portfólio. Uma breve descrição dessa técnica e do método de Markowitz é apresentada a seguir, após definirmos alguns conceitos fundamentais para o entendimento dos mesmos. Neste capítulo, apresentamos também definições e conceitos sobre *machine learning*, explicando a diferença entre os estudos supervisionado e não supervisionado, dando ênfase para a técnica de árvore de decisão.

No que segue, denotaremos por $\mathcal{P} = \{A_1, \ldots, A_N\}$ um portfólio qualquer, contendo os N ativos, A_i, $i = 1, \ldots, N$. Assumiremos que existe um ativo F, livre de risco, onde o investidor pode emprestar e tomar emprestado a uma taxa fixa R_F. O vetor dos pesos associados a esse portfólio $\mathcal{P}$ será denominado $\boldsymbol{w} = (w_1, ..., w_N)'$, onde w_i é peso atribuído ao ativo A_i, $i = 1, \ldots, N$. Os pesos são tais que $\sum_{i=1}^{N} w_i = 1$, ou seja, w_i representa a participação percentual do ativo A_i em relação ao total do portfólio. No decorrer deste trabalho, assumiremos que os pesos variam potencialmente com o tempo. Sendo assim, utiliza-se a notação $\boldsymbol{w}_t = (w_{1,t}, ..., w_{N,t})'$ para ressaltar esse fato e denotar o vetor de pesos correspondente ao tempo t.

[1] É o percentual de capital aplicado em cada ativo que compõe o portfólio.

2.1 Retornos e Agregação de Retornos

Em um portfólio, o risco é frequentemente medido em termos de variações de preços. No que segue, apresentamos as principais definições relacionadas à variação de preços de ativos financeiros, conforme descrito em Morettin e Toloi (2004); Morettin (2006).

Definição 2.1. (Retorno Líquido/Bruto Simples). Seja P_t o preço de um ativo no instante t. A variação de preços do ativo entre os instantes $t-1$ e t é dada por $\Delta P_t = P_t - P_{t-1}$ e a variação relativa de preços ou *retorno líquido simples* deste ativo entre os mesmos instantes é definida por

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{\Delta P_t}{P_{t-1}}.$$

Chamamos $1 + R_t = P_t/P_{t-1}$ de *retorno bruto simples.* Em geral, expressamos R_t em termos de porcentagem, relativamente ao período (um dia, um mês, um ano, etc). A quantidade R_t é também conhecida como *taxa de retorno.*

Definição 2.2. (Log-retorno). Definimos o *retorno composto continuamente* ou simplesmente *log-retorno* como

$$r_t = \ln(1 + R_t) = \ln(P_t) - \ln(P_{t-1}). \tag{2.1}$$

Da expressão (2.2) temos

$$1 + R_t = e^{r_t}, \quad \text{ou ainda,} \quad R_t = 1 - e^{r_t}.$$

Essas relações são utilizadas com frequência nas definições que seguem.

Definição 2.3. (Retorno de Período h). O *retorno simples de período h* entre os instantes $t-k$ e t é dado por

$$R_t[h] = \frac{P_t - P_{t-h}}{P_{t-h}} = \frac{P_t}{P_{t-h}} - 1. \tag{2.2}$$

A partir de (2.2), obtemos o *log-retorno de período h*, que é então definido por

$$r_t[h] = \ln(1 + R_t[h]) = \ln\left(\frac{P_t}{P_{t-h}}\right) = \sum_{j=0}^{h-1} r_{t-j}, \tag{2.3}$$

onde r_t é dado pela expressão (2.1) e o log-retorno de período 1 é r_t.

De acordo com Santos (2012), denotando por $R_{i,t+1}$ o retorno líquido do i-ésimo ativo e por $\boldsymbol{R}_{t+1} = (R_{1,t+1}, \cdots, R_{N,t+1})'$ o vetor de retornos aleatórios associado à

carteira com N ativos no tempo $t+1$, temos que o retorno da carteira de t a $t+1$ é dado por

$$R_{\mathcal{P},t+1} = \sum_{i=1}^{N} w_{i,t} R_{i,t+1}. \tag{2.4}$$

Nota-se que $R_{\mathcal{P},t+1}$ está condicionado aos respectivos pesos conhecidos no tempo t. A expressão (2.4) é denominada *agregação cross-section*.

Partindo de (2.4), conclui-se que o log-retorno da carteira de t a $t+1$ é dado por

$$r_{\mathcal{P},t+1} = \log(1 + R_{\mathcal{P},t+1}) = \ln\left(\sum_{i=1}^{N} w_{i,t}(1 + R_{i,t+1})\right) = \ln\left(\sum_{i=1}^{N} w_{i,t} e^{r_{i},t+1}\right),$$

onde utilizou-se o fato que $\sum_{i=1}^{N} w_{i,t} = 1$. Na prática, utiliza-se a aproximação (veja, por exemplo, Morettin, 2006)

$$r_{\mathcal{P},t+1} \approx \sum_{i=1}^{N} w_{i,t} r_{i,t+1}.$$

Sendo assim, sem perda de generalidade, no que segue, utilizamos as notações R_i $R_{i,t}$, $R_{\mathcal{P}}$ e $R_{\mathcal{P},t}$ para nos referirmos tanto a retornos quanto a log-retornos. Apenas nos casos em que se faz necessário, diremos explicitamente qual o tipo de retorno utilizado.

Alguns autores (por exemplo, Santos, 2012) optam ainda por trabalhar diretamente com o excesso de retorno, em vez do retorno em si. O **excesso de retorno** de um ativo, que é basicamente o valor que se consegue além do retorno em um investimento de baixo risco, é definido pela seguinte expressão:

$$R_i - R_F \tag{2.5}$$

onde R_i denota o retorno do i-ésimo ativo e R_F é o retorno do ativo livre de risco. O excesso de retorno do portfólio é definido de forma análoga. Como veremos na Seção 2.2, o modelo CAPM descreve a relação entre o valor esperado do excesso de retorno de um dado ativo e do mercado.

2.2 CAPM *(Capital Asset Pricing Model)*

Um das metodologias utilizadas para auxiliar na análise das características dos ativos que compõem o portfólio é o método CAPM. De acordo com (Tenani, 2016), o CAPM é o modelo de referência para a precificação de ativos financeiros, e pode ser definido como sendo a taxa de retorno que esperamos encontrar em um determinado ativo em relação a uma carteira diversificada e com risco controlado. O método

assume que os retornos do ativo e os retornos do mercado satisfazem a seguinte relação:

$$\mathbb{E}(R_i) = R_F + \beta_i[\mathbb{E}(R_m) - R_F], \quad i = 1, \ldots, N, \tag{2.6}$$

onde

R_i: retorno do ativo i
R_F: retorno livre de risco (Exemplos Brasil: renda fixa, poupança...)
β_i: risco associado ao investimento, usando a relação do retorno do ativo com o retorno do mercado
R_m: retorno do mercado

A versão empírica da equação (2.6) é a regressão linear, denominada *linha característica*, dada por

$$R_{i,t} - R_{F,t} = \alpha_i + \beta_i(R_{M,t} - R_{F,t}) + \varepsilon_{i,t}, \quad i = 1, \ldots, N \text{ e } t = 1, \ldots, n, \tag{2.7}$$

onde n é o tamanho amostral, $\varepsilon_{i,t}$ é o erro, representando o risco não sistemático.

O coeficiente β_i do modelo CAPM pode ser definido como a medida de volatilidade das taxas de retorno de um ativo com relação às taxas de retorno do mercado como um todo. Essa definição é motivada pelo fato de que, da maneira como o modelo CAPM é construído, β_i satisfaz a relação

$$\beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Var}(R_m)}. \tag{2.8}$$

O coeficiente angular β_i pode ser interpretado como (veja Prass, 2008, e referências ali contidas):

- $\beta_i = 1$. O ativo A_i pode ser considerado *neutro*, ou seja, à medida que os retornos do mercado sobem, o ativo sobe na mesma proporção, sendo a recíproca em relação à queda dos mercados também verdadeira.
- $\beta_i > 1$. O ativo A_i pode ser considerado *agressivo*, ou seja, à medida que os retornos do mercado sobem, o ativo sobe em uma proporção ainda maior, sendo a recíproca em relação à queda dos mercados também verdadeira.
- $\beta_i < 1$. O ativo A_i pode ser considerado *defensivo*, ou seja, à medida que os retornos do mercado sobem, o ativo também sobe, porém em uma proporção menor do que a que subiu o mercado, sendo a recíproca em relação à queda dos mercados também verdadeira.

De acordo com Prass (2008), utilizando a relação (2.4), mostra-se que o coefici-

ente β do portfólio $\mathcal{P}$, denotado por $\beta_{\mathcal{P}}$, pode ser expresso por

$$\beta_{\mathcal{P}} = \sum_{i=1}^{N} w_i \beta_i, \tag{2.9}$$

onde β_i, para $i = 1, \ldots, N$ é o coeficiente β do ativo A_i.

Conceitos paralelos ao CAPM, que auxiliam os investidores e analistas para embasar suas teses de investimentos, são o excesso de retorno, definido em (2.5), o α de Jensen e o índice Sharpe, cujas definições e interpretações são apresentadas a seguir.

Definição 2.4. (α de Jensen) O α de Jensen do ativo A_i, denotado por α_i, é definido como sendo o excesso de retorno do ativo menos o excesso de retorno esperado desse ativo,

$$\alpha_i = (R_i - R_F) - \mathbb{E}(R_i - R_F).$$

É fácil ver: como $\mathbb{E}(R_F) = R_F$, temos $\alpha_i = R_i - \mathbb{E}(R_i)$, ou seja, se o α_i for maior que zero, temos um retorno do ativo maior que o retorno esperado do ativo. Além disso, se o modelo CAPM dado em (2.6) for válido,

$$\alpha_i = (R_i - R_F) - [\mathbb{E}(R_i) - R_F] = (R_i - R_F) - \beta_i[\mathbb{E}(R_m) - R_F], \tag{2.10}$$

de forma que o valor esperado de α é zero. Na prática, uma estimativa para esse coeficiente é obtida através da regressão (2.7).

Definição 2.5. (Índice de Sharpe) O Índice de Sharpe de um ativo A_i, denotado por IS_i, é definido por

$$IS_i = \frac{R_i - R_F}{\sigma_i},$$

onde σ_i denota o desvio-padrão do retorno do ativo A_i, também chamado de risco do ativo.

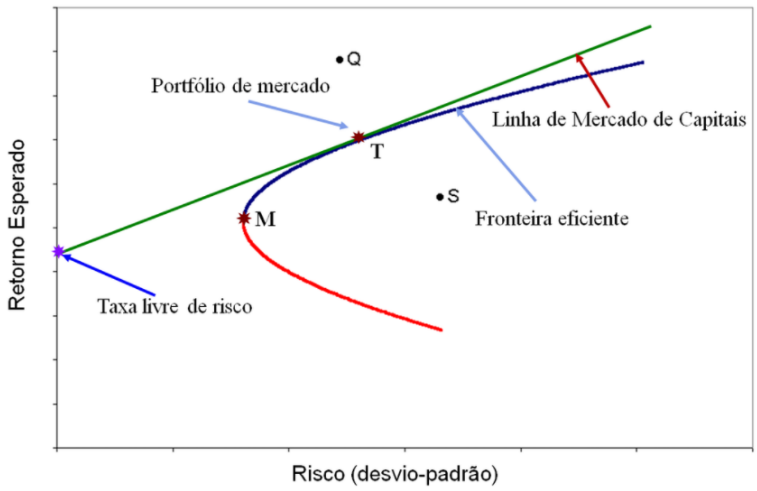
O excesso de retorno definido anteriormente está diretamente relacionado ao Índice de Sharpe, proposto por William Forsyth Sharpe (Sharpe, 1966), que nos mostra o quão bom são os investimentos realizados. Isso ocorre pelo fato de os termos utilizados nesse indicador levarem em conta o risco do ativo (medido pelo desvio-padrão), o retorno do ativo e o retorno livre de risco, os quais são medidas indispensáveis na hora de investir, visto que o investidor deseja maximizar os seus retornos em detrimento do risco. Sendo assim, quanto maior for o valor do Índice de Sharpe, melhores são nossas alocações.

Um ponto fraco dessa metodologia do CAPM é que ela se baseia em uma relação linear entre o retorno dos ativos e o retorno do mercado. Como apontado em Andriyashin et al. (2008), na prática, em geral essa relação é não linear e possui

uma forma paramétrica desconhecida, o que motiva o uso de modelos de *machine learning* para classificação dos ativos como neutro, agressivo ou defensivo.

2.3 Método Markowitz

O método de Markowitz propõe buscar uma alocação ótima dentro daquilo que ele chama de *fronteira eficiente*. O objetivo do método é que essa alocação, através da diversificação dos ativos, seja capaz de minimizar o risco da carteira em detrimento de um certo nível de retorno esperado. A Figura 2.1 representa a relação entre a fronteira eficiente, a linha de mercado de capitais (modelo CAPM), a taxa livre de risco (R_F), o portfólio do mercado, representado pela letra "T", e a carteira de mínima variância, representada pela letra "M".



Fonte: https://www.suno.com.br/artigos/fronteira-eficiente/

Figura 2.1: Fronteira Eficiente

As pessoas possuem graus diferentes de aversão ao risco, portanto, indivíduos que se arriscam mais, ou seja, aceitam uma maior volatilidade ou variância em seus portfólios, podem futuramente serem recompensados por um retorno também maior. Por outro lado, investidores mais conservadores procuram não se expor muito e buscam investimentos e carteiras com menor variância e, por conta disso, seus retornos são proporcionalmente menores também.

Existem dois tipos principais de operações realizadas na bolsa, com o objetivo de obter lucros: operar comprado e operar vendido. Operar comprado é realizar uma operação na qual o investidor obtém lucros com a valorização dos ativos; já operar vendido, ele lucra com a desvalorização deles. De maneira simplificada, podemos dizer que, em uma operação vendida, o investidor pega "emprestado" o ativo de uma outra pessoa que está alugando, então ele vai ao mercado e vende esses ativos. Passado um tempo, ele os recompra e "entrega" ao seu dono, pagando uma pequena

taxa pelo aluguel. Caso o ativo se desvalorize, o investidor ganha com essa diferença, pois o valor de compra terá sido menor que o vendido.

Conforme Santos (2012) traz em seu exemplo, podemos dizer que um investidor que pretenda operar apenas comprado tem a opção de diversificar os pesos que cada um desses ativos terá em sua carteira. No caso de operar comprado, assume-se que $w_i \geq 0$, para $1 \leq i \leq N$. Assumindo que o vetor $\boldsymbol{R}_t$ possui distribuição normal com média $\boldsymbol{\mu}_t = (\mu_{1,t}, ..., \mu_{N,t})$ e matriz de covariância $\Sigma_t = \{\sigma_{ij,t}\}$, da expressão (2.4), conclui-se que o retorno da carteira $R_{\mathcal{P},t} = \boldsymbol{w}_t'\boldsymbol{R_t}$ é normalmente distribuído, com média $\mu_{p,t} = \boldsymbol{w}_t'\boldsymbol{\mu}_t$ e variância $\sigma^2_{p,t} = \boldsymbol{w}_t'\Sigma_t\boldsymbol{w}_t$.

De acordo com Markowitz (1952), o desafio do investidor seria resolver o problema de minimização restrita, onde a carteira de média-variância é a solução do seguinte problema de otimização:

$$\underset{\boldsymbol{w}\in W}{\operatorname{argmin}}\left\{\boldsymbol{w}'\Sigma\boldsymbol{w} - \frac{1}{\gamma}\mathbb{E}(R_{\mathcal{P},t+1})\right\} \tag{2.11}$$

onde $W = \{\boldsymbol{w} \in \mathbb{R}^N, w_i \geq 0 \ \forall \ i = 1, ..., N, \text{ e } \sum_{i=1}^N w_i = 1\}$ e γ é nível relativo de aversão ao risco.

Um dos pontos fracos da metodologia tradicional de Markowitz é que ela, conforme mencionado anteriormente, se baseia na normalidade dos dados. E, além disso, envolve a estimação da matriz de covariância e a inversão da mesma, o que pode ser uma tarefa difícil, quando a quantidade de ativos envolvidos é muito grande. Levando-se em consideração as debilidades do método de Markowtiz, elencam-se os modelos de *machine learning*, como uma alternativa para flexibilizar a hipótese de normalidade e manipular grandes volumes de dados, como é o exemplo de Andriyashin et al. (2008), que utilizou árvore de decisão para construir portfólios em ativos da Bolsa de Valores da Alemanha.

2.4 *Machine Learning*

Com o fim do pregão viva voz na Bolsa de Valores, o aumento de pessoas físicas cadastradas na B3 e o uso de algoritmos que replicam estratégias de compra e venda de *traders* (nome dado às pessoas que se dedicam apenas a comprar ou vender as ações visando especular com a variação do preço), o volume de dados e de negociações aumentou e alcançou recordes históricos recorrentemente (como, por exemplo, Bona, 2019; Andrade, 2020). Por consequência, inúmeros pesquisadores, juntamente com investidores, têm buscado maneiras e métodos computacionais para auxiliar na tomada de decisão da compra ou venda de determinado ativo.

Como exemplos de investidores ou gestores que trabalham com os fundos chama-

dos quantitativos, ou seja, as decisões são tomadas por meio de algoritmos e modelos estatísticos, podemos citar a Leda Braga, que é uma brasileira CEO da *Systematica Investment* e tem sob gestão mais de USD 10 bilhões de dólares em ativos. Conforme informações na página ttps://www.suno.com.br/tudo-sobre/leda-braga/, ela obteve no fundo de investimento que gerenciava no ano de 2008 (ano de recessão econômica devido à crise do *subprime*) ganhos de mais de 40%, sendo que seus concorrentes tiveram uma média de retorno na casa dos 12%. Pode-se citar, também, Jim Simons, nascido nos Estados Unidos, formado em Matemática e com uma boa trajetória acadêmica, que decidiu utilizar seus conhecimentos matemáticos no mercado financeiro e fundou então a Renaissance Technologies, a qual contrata pessoas com conhecimentos sólidos em matemática, estatística, computação, entre outros. Vale destacar que a maioria das vagas não exige conhecimento financeiro prévio, uma vez que, conforme informado no próprio site da empresa (e.https://www.rentec.com/Home.action?index=true), ela é uma companhia que se dedica a produzir retornos excepcionais para seus investidores, aderindo estritamente a métodos matemáticos e estatísticos. Segundo informações na página https://www.suno.com.br/tudo-sobre/jim-simons/, o fundo da Renaissance é um dos fundos de investimentos mais rentáveis da história. Em 2021, Jim Simons tem, de acordo com a Forbes, uma fortuna pessoal de mais de USD 24 bilhões de dólares, o que o faz figurar entre as 100 pessoas mais afortunadas do mundo.

A previsão no mercado financeiro é um desafio relevante (Páscoa, 2018), haja vista que, encontrando boas estimações, é possível realizar ótimos e rentáveis negócios, sejam eles negócios individuais ou até mesmo formação de uma carteira profissional de investimento, como visto nos exemplos anteriores. Conforme Páscoa (2018), o desempenho preditivo é um dos benefícios mais reconhecidos de *Machine learning*.

De acordo com Mitchel (1997), formalmente, podemos definir *machine learning* como sendo um sistema computacional que busca realizar uma tarefa T, aprendendo a partir de uma experiência E, procurando melhorar uma performance P. Dentre as abordagens existentes, destacam-se o aprendizado supervisionado e o não supervisionado[2] (Burger, 2018).

No aprendizado não supervisionado, o objetivo é investigar a estrutura dos dados sem basear-se em um modelo com uma resposta previamente conhecida. Segundo James et al. (2013), o nome da abordagem se dá justamente pelo fato de não haver uma variável resposta para supervisionar a análise. Neste trabalho, focamos nossa atenção na abordagem supervisionada, descrita na seção que segue.

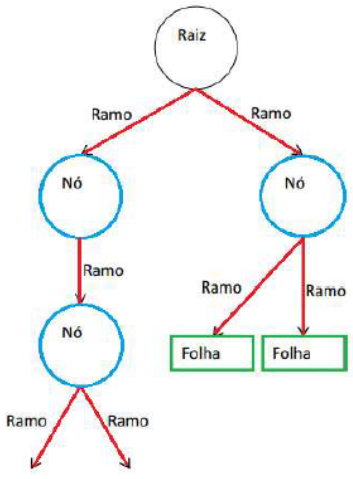
[2]Também existe a abordagem por reforço ou aprendizado "reinforçado" utilizados em outros cenários.

2.4.1 Aprendizado supervisionado

Uma das abordagens mais utilizadas em *machine learning*, de acordo com Burger (2018), é o aprendizado supervisionado. Essa metodologia é utilizada quando buscamos informações a partir de um conjunto de dados rotulados, ou seja, quando o conjunto de dados utilizado na fase de modelagem possui entradas (covariáveis) e saídas (variável resposta) conhecidas e o que se deseja construir (ou encontrar) é um modelo que seja capaz de aprender qual a melhor maneira de descrever/prever a resposta e chegar-se até ela, a partir das entradas fornecidas. Um dos modelos supervisionados que são simples e de fácil interpretação (James et al., 2013) é o modelo de árvore de decisão.

Árvore de decisão

Segundo Burger (2018), pode-se definir de maneira informal árvore como sendo uma estrutura que possui nós e arestas, onde para cada nó tem-se um valor que é dividido a fim de se obter informações dos dados. Na literatura de árvores de decisão, é comum também utilizarmos a nomenclatura de *nós*, *ramos* e *folhas* para designar os passos tomados, conforme mostra a Figura 2.2. Nesse contexto, os *nós* representam algum tipo de teste, os *ramos*, os resultados e as folhas são onde são especificados o que será retornado. Através de árvores de decisão, é possível descrever de forma gráfica e visual decisões a serem tomadas, eventos que podem vir a acontecer e, além disso, resultados combinados das decisões e eventos associados.



Fonte: Elaborada pelos autores

Figura 2.2: Árvore de decisão

Um exemplo simples trazido por Burger (2018) é a utilização do banco de dados denominado `mtcars`, disponível no R. A árvore de decisão associada ao problema é obtida com a utilização do código apresentado no quadro da Figura 2.3. Essa

figura também apresenta a árvore de decisão gerada pelo código. Nesse problema, a resposta é uma variável contínua (mpg = milhas por galão), e o objetivo é encontrar as principais características que influenciam a autonomia do veículo. Para os dados em questão, observou-se que o peso (wt) e a cilindrada (disp) são as variáveis que melhor descrevem a variável milhas por galão (mpg).

```
install.packages("party")

library(party)
arvore <- ctree(mpg ~ ., data = mtcars)
plot(arvore)
```

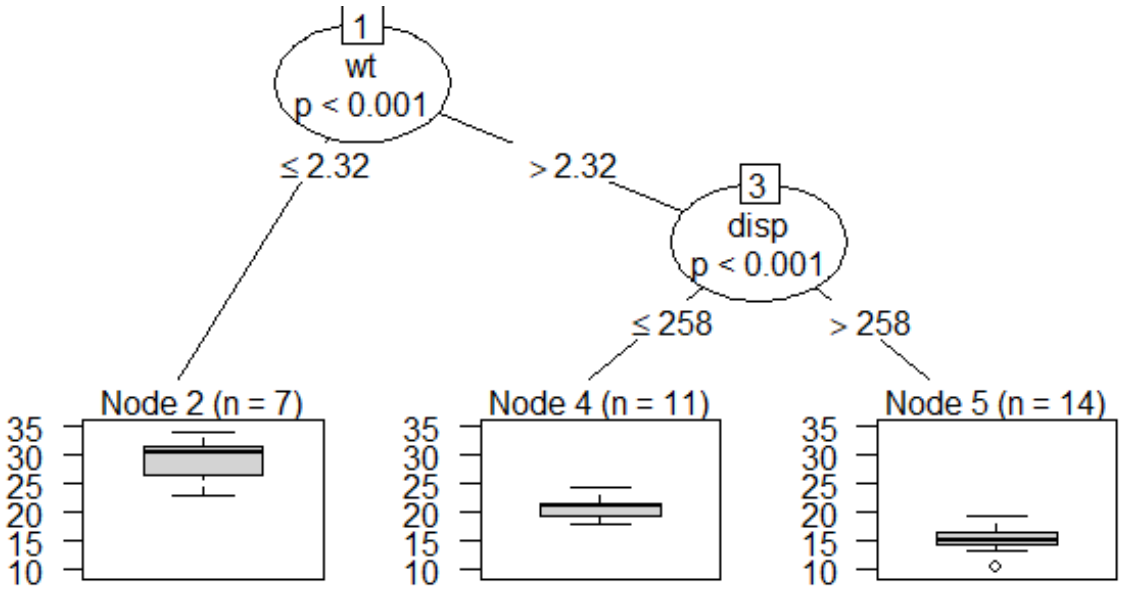


Figura 2.3: Árvore de decisão do banco mtcars

Existem diferentes algoritmos para construção das árvores de decisão, dependendo dos tipos das variáveis envolvidas. Dentre os algoritmos mais conhecidos, que servem como base para outros algoritmos mais complexos, podemos citar

- ID3 (*Iterative Dichotomiser 3*), proposto por Quinlan (1986), utilizado quando os atributos são categóricos e a resposta é binária.
- CHAID (*Chi-Squared Automatic Interaction Detector*), proposto por Kass (1980), utilizado quando todas as variáveis envolvidas são categóricas.
- CART (ou C&RT - *Classification and regression Tree*), proposto por Breiman et al. (1984), não possui restrições quanto ao tipo de variável.

No algoritmo ID3, inicia-se com todos os exemplos (observações) de treino em um mesmo grupo. A cada iteração do algoritmo, é escolhido o atributo (dentre os atributos que ainda não foram utilizados) que melhor divide os exemplos. Para cada atributo escolhido, cria-se um nó filho para cada valor possível do atributo e assim

repete-se o procedimento para cada filho não "puro"[3]. Existem três condições que forçam a parada do algoritmo: (a) todos os exemplos estão na mesma classe, (b) os exemplos não são da mesma classe, mas acabaram os atributos ou (c) não há exemplos no subconjunto (acontece quando nenhuma observação no conjunto pai corresponde a um valor específico do atributo selecionado). Basicamente, a ideia do *algoritmo ID3* é pôr à prova os atributos mais importantes primeiro pois, desta forma, espera-se conseguir a classificação correta com um pequeno número de nós (testes).

O algoritmo CHAID tem por base o teste qui-quadrado em uma tabela de contingência entre as categorias da variável dependente e as categorias das variáveis independentes. Caso existam variáveis contínuas no banco, elas devem, primeiramente, ser discretizadas em classes. Após a categorização das variáveis, é realizada uma série de testes para encontrar o melhor número de classes da variável de entrada; depois é encontrada a melhor variável explicativa e, por fim, decide-se se é plausível a realização de uma divisão adicional sobre o nó. O Exhaustive CHAID (Bigss et al., 1991) é uma modificação do algoritmo CHAID. Nele, o processo de juntar categorias e testar variáveis preditoras é mais completo e, portanto, requer maior tempo computacional.

Neste trabalho, utilizamos a versão do algoritmo CART implementada no pacote `rpart` do R (Therneau e Atkinson, 2019). Nesse pacote, o algoritmo recebe o nome de RPART (*Recursive Partitioning And Regression Trees*). Segundo os autores do pacote, o uso do termo RPART deve-se ao fato de o nome CART ser uma marca registrada de um *software* que possui uma implementação específica das ideias do algoritmo. O algoritmo CART segue passos semelhantes aos algoritmos anteriores. O primeiro passo é encontrar uma variável que melhor divide os dados em dois grupos (onde "melhor" é definido em termos de alguma medida de qualidade de ajuste). Após a separação dos dados, ocorre separadamente em cada subgrupo novamente esse processo de maneira recursiva, até que nenhuma melhoria possa ser feita. Os critérios adotados para separação e avaliação da qualidade de separação dependem do tipo de problema, se classificação ou regressão. Após esse estágio, ocorre a etapa de poda da árvore, onde utiliza-se validação cruzada para determinar o número de nós terminais ideal, com base em um critério envolvendo um parâmetro de complexidade α. Uma descrição completa desse algoritmo pode ser encontrada no manual fornecido pelos autores do pacote `rpart`, no link `https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf`.

Conforme Lauretto (2010), o processo de divisão que seria o "melhor" seria o que agrupasse da forma mais adequada possível os exemplos de mesma classe. Por

[3]Um filho puro é aquele em que cada atributo específico tem o mesmo valor em todos os exemplos.

outro lado, após a expansão da árvore, pode ocorrer que os subconjuntos formados nos nós e nas folhas sejam muito pequenos e com grande efeito de *overfitting* (ajuste muito preciso aos dados de treinamento, porém com baixa precisão na classificação de novos exemplos). Sendo assim, a poda é útil para eliminar esses ramos e serve para manter na árvore apenas as regras com alto grau de distinção das classes.

3 Estudo Empírico

Grande parte dos investidores realiza seus investimentos baseados nos fundamentos da empresa, ou seja, fluxo de caixa, aumento do lucro, distribuição dos dividendos, indicadores macroeconômicos, entre outros. Cabe destacar que, no entanto, a decisão final baseada nesses critérios pode sofrer de parcialidade, visto que a "última palavra" acaba sendo do investidor ou gestor.

Por exemplo, digamos que um gestor de um fundo aprecie empresas que nos últimos 10 anos tenham apresentado lucros seguidos e distribuição de dividendos, e que ele só invista em empresas que tenham essas duas características[1] conjuntamente, e que, ano após ano, ele realiza um rebalanceamento em sua carteira, comprando ou vendendo os ativos que, de acordo com a "regra"[2] de alocação, devem permanecer no fundo. Entretanto, no ano de 2019, a empresa, que daremos o nome de *empresa genérica*, não apresentou lucro, porém distribuiu dividendos. Sendo assim, pela "regra" ou "filosofia" de investimentos do fundo, o gestor não poderia permanecer com o ativo para o próximo ano; todavia, a *empresa genérica* conta com o maior percentual de alocação no fundo e está presente no mesmo desde os anos 1990, sem contar que o noticiário aparenta estar favorável e várias casas de análises estão recomendando a compra. O gestor, então, "quebra" a sua própria regra e continua com a *empresa genérica*.

O fato de o gestor continuar investindo na *empresa genérica* é causado por uma parcialidade, gosto pessoal e não pela sua filosofia de investimentos, e a quebra dessa filosofia pode parecer ingênua no presente, porém no longo prazo, ter uma boa métrica de investimentos e segui-la é fundamental. No livro *Investindo em ações para o longo prazo*, o autor Siegel (2015) alerta sobre a importância de se ter regras "firmes", para manter a carteira de investimentos nos "trilhos", principalmente na ocasião em que o gestor ou investidor sente que está cedendo à emoção do momento,

[1] É apenas um exemplo didático, para evidenciar o possível viés que o gestor pode ter ao selecionar determinado ativo.

[2] Foi colocado aspas em regra, pois não tem a ver com as leis e regulamentações da CVM (Comissão de Valores Mobiliários) que regem o fundo, como as informações contidas no prospecto, ou lâmina, e sim com a filosofia empírica do mesmo.

como foi o caso em que vimos anteriormente da permanência da *empresa genérica* no portfólio.

Após esse preâmbulo, é reforçada a ideia de seguir as chamadas "regras firmes" conforme Siegel (2015), sejam elas do ponto de vista em análises fundamentalistas, técnicas, quantitativas, entre outras. Por isso, no presente trabalho, optou-se, então, pelas árvores de decisão, pois nelas, executamos um algorítmico para a seleção de ativos da Bolsa de Valores brasileira, em que são definidas regras iniciais que, posteriormente, devem ser seguidas para se ter um retorno esperado, visto que, até o momento, tem-se poucos estudos utilizando a metodologia de árvores de decisão na B3. A fim de validar e verificar se a escolha do algorítmico de árvore de decisão teve um bom retorno, foram comparados os seus resultados com outras métricas de otimização de portfólio e indicadores do mercado, que elucidaremos nas próximas seções.

3.1 Dados utilizados

Neste trabalho, foram utilizadas séries históricas disponíveis nos sites do Yahoo Finanças (Yahoo, 2020), B3 (B3, 2020) e Banco Central do Brasil (Selic, 2020). O banco de dados que foi empregado na pesquisa consiste nos log-retornos dos preços ajustados de ações comercializadas na Bolsa de Valores brasileira B3, portanto, dados reais.

Os ativos que foram escolhidos para compor o portfólio são apenas os ativos que pertencem ao Índice Bovespa, uma vez que possuem mais liquidez e são mais fáceis de utilizar com os pacotes usados ao longo da pesquisa. Porém, cabe destacar que a metodologia utilizada neste trabalho pode ser replicada para qualquer outro grupo de ativos e qualquer período de escolha, desde que haja um número suficiente de observações para que os métodos possam ser empregados.

3.1.1 Período selecionado

Ao longo da pesquisa, optou-se por utilizarmos os dados históricos de 01/01/2012 até 31/12/2019, visto que este período contempla fases de lateralização[3] do Índice Bovespa (IBOV) e de algumas tendências, tanto de alta quanto de baixa. Os dados referentes ao período de 01/01/2012 a 31/12/2018 foram utilizados para o treinamento das árvores de decisão e estimação dos pesos dos portfólios referentes ao mês de 01/01/2019 (em todos os modelos considerados). Já os dados referentes ao

[3]Dizemos que um ativo está lateralizado quando não apresenta tendência de alta e nem baixa, ou seja, permanece estável ao longo do tempo.

período de 01/01/2019 a 01/12/2019, aqui denominados base de validação, foram utilizados para atualização dos pesos dos portfólios e comparação dos rendimentos dos diferentes modelos. Mais detalhes sobre a utilização dos dados em cada contexto são fornecidos a seguir.

3.1.2 Pré-processamento dos dados

Tendo em vista que o trabalho não tinha por objetivo explorar as técnicas em questão na presença de dados faltantes, os ativos do IBOV que não tiveram série histórica completa durante os anos selecionados de 2012 até 2019 foram removidos da análise. Ao final desta etapa restaram 67 ativos, cujas siglas são listadas na Tabela 3.1.

Tabela 3.1: Ativos utilizados na pesquisa

ABEV3.SA	B3SA3.SA	BBAS3.SA	BBDC3.SA
BBDC4.SA	BEEF3.SA	BRAP4.SA	BRFS3.SA
BRKM5.SA	BRML3.SA	BTOW3.SA	CCRO3.SA
CIEL3.SA	CMIG4.SA	CPFE3.SA	CPLE6.SA
CSAN3.SA	CSNA3.SA	CYRE3.SA	ECOR3.SA
EGIE3.SA	ELET3.SA	ELET6.SA	EMBR3.SA
ENBR3.SA	ENEV3.SA	ENGI11.SA	EQTL3.SA
EZTC3.SA	FLRY3.SA	GGBR4.SA	GOAU4.SA
GOLL4.SA	HGTX3.SA	HYPE3.SA	IGTA3.SA
ITSA4.SA	ITUB4.SA	JBSS3.SA	JHSF3.SA
LAME4.SA	LREN3.SA	MGLU3.SA	MRFG3.SA
MRVE3.SA	MULT3.SA	PCAR3.SA	PETR3.SA
PETR4.SA	PRIO3.SA	QUAL3.SA	RADL3.SA
RENT3.SA	SANB11.SA	SBSP3.SA	SULA11.SA
SUZB3.SA	TAEE11.SA	TIMS3.SA	TOTS3.SA
UGPA3.SA	USIM5.SA	VALE3.SA	VIVT3.SA
VVAR3.SA	WEGE3.SA	YDUQ3.SA	

3.2 Indicadores comparáveis

No que segue, descrevemos com mais detalhes os dois indicadores utilizados para comparação, bem como os passos adotados para cada um dos métodos considerados no cálculo dos pesos do portfólio.

3.2.1 Taxa Selic

Escolher uma taxa livre de risco é basicamente encontrar um investimento que possua um risco bastante reduzido, ou seja, a chance de o valor investido se perder tende a ser pequena. Contudo, por a taxa livre de risco tender a ter pouca volatilidade, normalmente o seu retorno também é reduzido. No Brasil, conforme Bona (2020), costuma-se utilizar investimentos como o Tesouro Selic, que são aplicações de baixo risco, nas quais pode-se avaliar o custo de oportunidade.

No presente trabalho, utilizaremos a taxa Selic, cuja sigla vem de: *Sistema Especial de Liquidação e de Custódia*. Conforme BCB (2021), a Selic é a taxa básica de juros da economia brasileira, sendo responsável por influenciar várias taxas de juros no Brasil, como as taxas de juros dos empréstimos, dos financiamentos e das aplicações financeiras.

Esse foi um dos indicadores que comparamos com a seleção de portfólio via árvores de decisão. O código para importação dos dados da Selic baseou-se no código implementado por Araujo JR (2018), sendo feitas alterações pontuais para adaptar ao cenário do trabalho.

3.2.2 Índice Bovespa

O Ibovespa é um dos indicadores utilizados para avaliar o desempenho do mercado acionário brasileiro. Acontece a cada quatro meses um rebalanceamento, haja vista que o índice é oriundo de uma carteira teórica de ativos conforme explicado pela própria B3 (2021).

De acordo com o que foi elucidado anteriormente, o período escolhido para análise foi de 2012 até 2019, porém alguns ativos acabaram não tendo série histórica suficiente e saíram da análise. Para o Ibovespa, foi considerado apenas o seu valor de preço ajustado e assim calculado o retorno ao longo desse período. Não verificamos se os ativos removidos e/ou utilizados nas demais análises entraram ou não na composição do Ibovespa ao longo do período analisado.

3.2.3 Método *Naive*

A carteira *Naive*, também chamada de carteira ingênua, é aquela que distribui o peso[4] de maneira igual para todos os ativos, em qualquer tempo. Sendo assim, o

[4] É o percentual de capital aplicado em cada ativo que compõe o portfólio.

peso $w_{i,t}$, correspondente ao ativo A_i, no tempo t, é dado por

$$w_{i,t} = \frac{1}{N}, \quad i = 1, \ldots, N, \text{ e } t > 0, \tag{3.1}$$

onde N é o número de ativos no portfólio. Esse foi um dos métodos que comparamos com a seleção de portfólio via árvore de decisão.

Ressaltamos que, para o método *Naive*, a divisão dos dados em período de treinamento e validação/teste não interfere na estimação dos pesos. Os dados referentes ao período de treinamento não foram utilizados em nenhum momento. Os dados referentes aos período de validação (12 meses) foram utilizados para calcular os retornos do portfólio formado pelos 67 ativos em questão, de forma que $w_{i,t} = 1/67$.

3.2.4 Método de Markowitz

Os dados do período de treinamento foram utilizados para estimar os pesos ótimos correspondentes a jan/2019. Em seguida, os dados do período de validação foram incorporados, mês a mês, para estimação dos pesos correspondente ao mês seguinte. Por exemplo, utilizaram-se os dados do início de 2012 até 31/01/2019, para estimar os pesos correspondentes a fev/2019 e assim sucessivamente, visto que foi utilizada uma janela variável de tamanho do início até o novo mês. Vale destacar que, com os dados disponíveis, é possível estimar 13 carteiras, porém apenas as 12 primeiras são utilizadas, uma vez que os dados referentes ao último mês 2019 podem ser utilizados para atualização dos pesos, porém não temos dados após essa data para cálculo de retornos, para fins de comparação.

Os pesos ótimos foram estimados através da função `globalMin.portfolio` disponível no pacote `IntroCompFinR` (Zivot, 2015). Para o cálculo da carteira de mínima variância, definiu-se como falso (`FALSE`) o argumento `shorts`, tendo em vista que os ativos foram operados apenas de maneira comprada e não ocorreu a venda deles. A função em questão resolve o problema de otimização, que corresponde a encontrar o mínimo de $\boldsymbol{w}'\Sigma\boldsymbol{w}$, restrito a $\sum_{i=1}^{N} w_{i,t} = 1$.

No total, 25 ativos distintos foram selecionados ao longo de 12 meses. A composição da carteira para cada mês de 2019 é apresentada na Tabela 3.2. Nessa tabela, o símbolo "-" indica que o ativo em questão não participou do portfólio naquele mês. Devido à magnitude dos valores, multiplicamos os pesos por 100 para auxiliar na interpretação.

Tabela 3.2: Pesos via Markowitz utilizados na pesquisa

	Mês											
Ativo	1	2	3	4	5	6	7	8	9	10	11	12
SUZB3.SA	23,95	19,88	19,89	19,41	19,32	15,31	15,04	15,67	15,49	14,27	14,26	12,80
WEGE3.SA	10,87	14,26	14,47	14,50	14,44	17,09	17,10	16,71	14,40	13,97	14,44	11,51
EMBR3.SA	9,98	13,04	12,85	12,73	12,80	12,71	12,59	12,62	12,58	13,05	12,98	14,14
VIVT3.SA	0,50	3,65	4,41	5,64	5,61	8,92	9,11	8,30	9,84	11,69	11,66	11,19
PCAR3.SA	4,91	5,61	5,76	5,83	5,84	5,85	5,80	5,90	6,48	7,07	7,01	6,89
MULT3.SA	7,17	7,84	7,68	6,59	6,59	6,25	5,75	6,01	5,20	3,54	3,04	5,80
UGPA3.SA	4,62	8,90	8,05	6,99	6,81	7,00	7,61	8,45	5,35	3,16	3,47	0,30
EGIE3.SA	9,44	1,18	1,32	2,89	2,99	5,30	6,36	5,45	3,02	4,37	4,53	3,97
EQTL3.SA	1,78	2,77	1,98	1,67	1,57	3,17	3,19	3,72	4,61	5,34	5,28	5,66
RADL3.SA	1,87	2,65	3,13	3,28	3,35	2,84	2,56	2,55	3,64	4,25	4,56	5,42
BRKM5.SA	2,73	3,46	3,49	3,10	3,11	2,74	3,00	3,14	2,68	2,56	2,36	3,30
CIEL3.SA	8,46	3,76	3,58	3,00	2,89	1,00	0,76	0,44	2,24	2,52	2,21	3,58
BRFS3.SA	3,46	3,63	2,88	3,51	3,75	1,52	1,35	1,23	2,37	2,65	2,28	3,82
BEEF3.SA	4,47	1,69	2,49	2,82	2,82	3,28	3,23	3,11	2,49	2,68	2,91	-
MRFG3.SA	0,98	2,31	1,89	1,90	1,91	1,92	2,15	2,24	2,60	1,54	1,53	2,20
VALE3.SA	-	2,06	1,97	2,06	2,01	2,02	2,03	2,28	1,73	1,62	1,64	1,47
HGTX3.SA	3,19	2,07	2,18	1,85	1,79	0,86	0,81	0,83	1,33	1,13	1,03	1,21
CPFE3.SA	-	-	-	-	-	0,18	0,29	0,23	1,11	1,46	1,46	3,23
CYRE3.SA	1,00	0,89	1,62	1,87	2,03	0,54	-	-	-	0,26	0,90	-
ENEV3.SA	0,63	0,34	0,34	0,36	0,37	0,74	0,64	0,57	0,79	0,62	0,61	0,68
ENGI11.SA	-	-	-	-	-	0,58	0,64	0,54	0,62	-	-	2,42
SULA11.SA	-	-	-	-	-	0,18	-	-	1,08	1,27	1,29	0,39
RENT3.SA	-	-	-	-	-	-	-	-	-	0,94	0,53	0,02
QUAL3.SA	-	-	-	-	-	-	-	-	0,36	-	-	-
ENBR3.SA	-	-	-	-	-	-	-	-	-	0,01	-	-

3.3 Árvore de decisão

Nesta seção, descrevemos os passos adotados para construção de um portfólio, utilizando árvores de decisão. A metodologia proposta foi inspirada em uma aplicação apresentada por Guerra (2018). Guerra considera os valores de abertura (*open*), fechamento (*close*), mínimo (*low*) e máximo (*high*) do Ibovespa e algumas variáveis auxiliares definidas a partir dessas quatro e seus valores defasados (do tempo $t-1$ até, no máximo, o tempo $t-10$). Para o autor, a variável resposta (*target*) é uma variável binária, a qual assume valor 1 se a diferença entre o máximo e a abertura for maior ou igual do que 500, e zero, caso contrário. Um modelo de árvore de decisão foi utilizado para prever a resposta um passo à frente.

Neste trabalho, para cada ativo A no banco de dados, definimos a seguinte variável (*target*):

$$I_A(t) = \begin{cases} 1, & \text{se} \quad (R_{A,t} - R_{F,t})/|R_{F,t}| \geq 0,10 \\ 0, & \text{caso contrário,} \end{cases}$$

onde $R_{F,t}$ é o retorno da Selic e $R_{A,t}$ é o retorno do ativo A, no tempo t. Note que $I_A(t)$ indica se o retorno do ativo foi maior do que o da Selic em pelo menos 10%. Dois modelos de árvore de decisão foram propostos para prever o *target*. O primeiro deles (M1) considera como covariável o log-retorno do ativo A no tempo $t-1$ e o segundo (M2), considera os log-retornos do ativo nos tempos $t-1$ e $t-2$.

As árvores de decisão foram obtidas com a função `rpart`, disponível no pacote de mesmo nome (Therneau e Atkinson, 2019). Para o ajuste dos modelos, consideraram-se os dados do período de 01/01/2012 a 31/12/2018. Esses dados foram divididos em duas partes, de forma aleatória: treinamento (80% dos dados) e teste (20% dos dados). Os dados de treinamento foram utilizados para treinar as árvores de decisão. Após obter as duas árvores (M1 e M2), realizou-se a previsão para os dados de teste. A árvore que obteve o melhor desempenho nessa etapa foi selecionada e definida como sendo a árvore de decisão associada ao ativo em questão. Todas as previsões foram obtidas utilizando-se a função `predict`, que internamente invoca a função `predict.rpart`, também do pacote `rpart`.

Para decidir quais ativos devem compor o portfólio em um determinado instante de tempo t e quais devem ser seus pesos, definiu-se a seguinte variável:

$$S_A(t) = \sum_{j=1}^{t} \hat{I}_A(j),$$

onde $\hat{I}_A(j)$ denota a previsão do *target* $I_A(j)$, no tempo j. Essa variável nada mais é do que a previsão para o número de vezes que o retorno do ativo é pelo menos 10% maior do que o retorno da Selic do início do período até o tempo t. A partir de $S_A(\cdot)$, define-se o seguinte:

- **Critério de seleção.** Serão selecionados para compor o portfólio no tempo t todos os ativos que tiveram as duas maiores pontuações previstas para o tempo em questão, isto é, os maiores valores de $S_A(t)$, para evitar que apenas um ativo fique em carteira.

- **Pesos**. Supondo que k ativos $A_1, \ldots, A_k$ sejam selecionados para compor o portfólio no tempo t, o peso atribuído a cada um deles é

$$w_{i,t} = \frac{S_{A_i}(t)}{\sum_{j=1}^{k} S_{A_j}(t)}. \quad i = 1, \ldots, k.$$

Após selecionar a árvore de decisão associada a cada ativo, obtiveram-se todas as previsões $\{\hat{I}_A(t)\}_{t=1}^{n}$ das variáveis *target* para o período de 2012 a 2018, onde $n = 83$ é o número de observações mensais nesse período. Depois de obter os valores de $S_A(n)$ para os 67 ativos do banco de dados, utilizou-se o critério descrito anteriormente para definir o portfólio referente a jan/2019 (tempo $t = n + 1 = 84$). Para os meses subsequentes, repetiu-se o procedimento de previsão, utilizando as árvores previamente selecionadas. Ressaltamos que, nessa etapa, utilizamos os valores dos log-retornos que não participaram do treinamento da árvore, exceto na construção da carteira de fev/2019, em que utilizamos os log-retornos de jan/2019 (que não havia sido utilizado no treinamento da árvore) e dez/2018 (sendo esse último utilizado apenas quando M2 foi selecionada). É fácil ver que

$$S_A(t+1) = S_A(t) + \hat{I}_A(t+1)$$

ou ainda,

$$S_A(t+h) = S_A(t) + \sum_{j=1}^{h} \hat{I}_A(t+j), \quad h > 0,$$

ou seja, os novos valores da variável acumulada podem ser obtidos somando-se as novas previsões aos valores anteriores da variável acumulada.

No total, 7 ativos distintos foram selecionados ao longo de 12 meses. A composição da carteira para cada mês de 2019 é apresentada na Tabela 3.3. Nessa tabela, o símbolo "-" indica que o ativo em questão não participou do portfólio naquele mês. Devido à magnitude dos valores, multiplicamos os pesos por 100 para auxiliar na interpretação.

Tabela 3.3: Pesos via árvore de decisão utilizados na pesquisa

	Mês											
Ativo	1	2	3	4	5	6	7	8	9	10	11	12
ENGI11.SA	32,00	32,03	32,03	32,03	47,17	47,22	47,27	47,32	46,90	46,96	19,42	13,95
EQTL3.SA	36,00	35,95	35,95	35,95	52,83	52,78	52,73	52,68	53,10	53,04	22,30	16,28
IGTA3.SA	32,00	32,03	32,03	32,03	-	-	-	-	-	-	19,42	13,95
HYPE3.SA	-	-	-	-	-	-	-	-	-	-	19,42	13,95
MGLU3.SA	-	-	-	-	-	-	-	-	-	-	19,42	13,95
QUAL3.SA	-	-	-	-	-	-	-	-	-	-	-	13,95
RADL3.SA	-	-	-	-	-	-	-	-	-	-	-	13,95

3.4 Resultados

Para cada método considerado neste trabalho, calculou-se o retorno que seria obtido pelo portfólio, caso os pesos em questão fossem adotados. Os resultados obtidos são apresentados na Tabela 3.4, para o período de jan/2019 a dez/2019. Nessa tabela, apresentamos ainda o log-retorno da Selic e do Ibovespa, no mesmo período. Além dos resultados mensais das cinco abordagens de investimentos, também na Tabela 3.4, são exibidos o retorno médio, desvio-padrão, retorno acumulado, índice de Sharpe e o *Turnover*, que foi calculado utilizando a seguinte equação:

$$TO_t = \sum_{i=1}^{N} \left| w_{i,t+1} - w_{i,t} \frac{1 + R_{i,t}}{1 + \boldsymbol{w}_t' \boldsymbol{R}_t} \right| \tag{3.2}$$

a qual, basicamente, mede a oscilação dos pesos dos ativos presentes na carteira de investimentos. Portanto, podemos inferir que quanto maior for o valor do *turnover*, maiores serão os custos de operação.

Tabela 3.4: Retorno dos métodos e indicadores

Data	Selic	Ibovespa	Naive	Markowitz	Árvores
31/01/2019	0,0053	0,1027	0,1284	0,1277	0,1199
28/02/2019	0,0048	-0,0188	-0,0185	-0,0134	-0,0509
29/03/2019	0,0046	-0,0018	-0,0042	-0,0248	-0,0304
30/04/2019	0,0051	0,0098	0,0198	0,0036	0,0205
31/05/2019	0,0053	0,007	0,0186	-0,057	0,0665
28/06/2019	0,0046	0,0398	0,0438	0,028	0,0438
31/07/2019	0,0056	0,0083	0,0579	0,0325	0,0490
30/08/2019	0,0049	-0,0067	0,0132	-0,0441	-0,0114
30/09/2019	0,0045	0,0351	0,0313	0,0573	0,0490
31/10/2019	0,0047	0,0234	0,0147	-0,0052	-0,0088
29/11/2019	0,0037	0,0094	0,0493	0,0836	-0,0120
30/12/2019	0,0036	0,0662	0,0926	0,0715	0,0733
Retorno Médio	0,0048	0,0228	0,0372	0,0216	0,0257
Desvio-Padrão	0,0006	0,0340	0,0411	0,0555	0,0497
Retorno Acumulado	0,0578	0,2745	0,4470	0,2597	0,3086
Índice de Sharpe	——	0,5299	0,7877	0,3018	0,4209
Turnover	——	——	0,0608	0,1676	0,2267

Observa-se que a metodologia que utiliza árvores de decisão obteve o segundo maior retorno dentre os cinco no período selecionado. É válido destacar que o método *Naive*, mesmo sendo uma alocação ingênua, ainda assim obteve um retorno bem acima do Markowitz, índice do mercado, árvores de decisão e, inclusive, um valor para o *turnover* menor dentre os três métodos avaliados, porém teve o maior

Índice de Sharpe dentre as abordagens, o que advém de o método *naive* ter incorrido uma alta volatilidade frente aos demais, para entregar um bom retorno acumulado.

Apenas para fins de comparação, caso fossem investidos R$ 1000,00 em cada uma das cinco abordagens, teríamos aproximadamente o seguinte retorno:

Tabela 3.5: Retorno linear do período de jan/2019 a dez/2019

-	Selic	Ibovespa	*Naive*	Markowitz	Árvores
%	5,95	31,58	56,36	29,65	36,15
R$1000,00	R$1059,46	R$1315,84	R$1563,60	R$1296,50	R$1361,47

4 Conclusão

O presente trabalho tinha como finalidade explorar e aplicar técnicas de *Machine Learning*, em especial as árvores de decisão para conseguir selecionar um portfólio de ações da Bolsa de Valores do Brasil. No que tange aos resultados das cinco abordagens: árvores de decisão, *Naive*, Ibovespa, Markowitz, Selic, é válido enfatizar a boa performance das árvores de decisão e da alocação ingênua, evidenciando um bom ajuste para o período selecionado quando comparado aos retornos do mercado e o próprio Markowitz, que, embora tenham sido positivos, ficaram abaixo dos dois primeiros. Em relação à taxa livre de risco usada, cabe lembrar que, a partir do ano de 2016, a Selic passou por uma contínua redução da taxa de juros, o que acabou tendo um impacto significativo no retorno da mesma, visto que o período do estudo foi de 2012 até 2019.

Quanto a novas ideias na abordagem de seleção de portfólio de ações utilizando árvores de decisão, pode-se sugerir outros ativos para criação da base de dados pois, em vez de ficar restrito apenas aos ativos do IBOV, é razoável selecionar outros papéis listados na Bolsa que tenham liquidez diária suficiente para haver negócios. Testar o método para os Fundos Imobiliários que apresentam boa liquidez pode ser tão promissor quanto foi para as ações. Avaliar o desempenho das árvores alterando o *target* pode apresentar resultados diferentes aos anteriores, sendo que o efeito pode ser tanto positivo quanto negativo.

Concluímos, portanto, que há muito espaço para novas ideias e proposições de métodos matemáticos, estatísticos e computacionais para serem estudados não só no mercado mundial, mas principalmente no mercado brasileiro, que está em constante expansão. Conforme afirmou Siegel (2015), o mercado acionário é algo empolgante, porém podemos dizer que estudá-lo torna-o ainda mais.

Referências Bibliográficas

Andrade, J. (2020). B3 bate recorde e movimenta R$ 26 bilhões por dia em 2020. https://einvestidor.estadao.com.br/investimentos/b3-recorde-26-bilhoes-dia/.

Andriyashin, A., Härdle, W., e Timofeev, R. (2008). Recursive portfolio selection with decision trees. *SFB 649 Discussion Paper Economic Risk Berlin*, 9:1–27.

Araujo JR, B. J. (2018). Importação da Taxa Selic. https://rpubs.com/jbajr1982/392718.

B3 (2020). B3: A Bolsa do Brasil. http://www.b3.com.br.

B3 (2021). Ibovespa B3. http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm.

BCB (2021). Taxa Selic. https://https://www.bcb.gov.br/controleinflacao/taxaselic.

Bigss, D., Ville, B., e Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. *Journal of Applied Statistics*, 18(1):49–62.

Bona, A. (2019). B3 registra maior volume financeiro de negociação de sua história. https://andrebona.com.br/b3-registra-maior-volume-financeiro-de-negociacao-de-sua-historia/.

Bona, A. (2020). Taxa livre de risco e prêmio de risco: entenda mais sobre esses conceitos. https://tinyurl.com/h2xcn7xt.

Breiman, L., Friedman, J., Olshen, R., e Stone., C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Burger, S. V. (2018). *Introduction to Machine Learning with R*. O'Reilly.

Fortuna, E. (2007). *Mercado Financeiro: produtos e serviços*. Qualitymark.

Guerra, L. (2018). Aplicando árvores de decisão no Ibovespa - Finanças Quantitativas. https://www.outspokenmarket.com/blog/aplicando-arvores-de-decisao-no-ibovespa-financas-quantitativas.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction To Statistical Learning*. Springer.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 20(2):119–127.

Lauretto, S. M. (2010). Árvores de Decisão. https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresDecisao_normalsize.pdf.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Mitchel, T. (1997). *Machine Learning*. McGraw-Hill.

Morettin, P. (2006). *Um Curso em Séries Temporais Financeiras. Minicurso do 17º Sinape*. São Paulo: ABE.

Morettin, P. e Toloi, C. (2004). *Análise de Séries Temporais*. São Paulo: Edgard Blücher.

Prass, T. S. (2008). *Análise e Estimação de Medidas de Risco em Processos FIEGARCH*. PhD thesis, Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

Prass, T. S. e Lopes, S. R. (2012). Var, teste de estresse e maxloss na presença de heteroscedasticidade e longa dependência na volatilidade. *Revista Brasileira de Estatística*, 73(5):47–80.

Páscoa, M. I. F. (2018). *Os desafios da Machine Learning: Aplicação ao Mercado Financeiro*. Faculdade de Economia Universidade de Coimbra.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1):81–106.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Santos, André Alves Portela; Tessari, C. (2012). Técnicas quantitativas de otimização de carteiras aplicadas ao mercado brasileiro de ações. *Revista Brasileira de Finanças*, 13(27089):369–393.

Selic (2020). Selic BACEN. https://www.bcb.gov.br.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, 39(1):119–138.

Siegel, J. J. (2015). Investindo em ações no longo prazo . 5ª Edição.

Tenani, P. S. (2016). Revisitando o CAPM . Disponível em GV Invest 03 Short Studies Series EESP-FGV.

Therneau, T. e Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Yahoo (2020). Yahoo Finanças. https://br.financas.yahoo.com/.

Zivot, E. (2015). *IntroCompFinR: Introduction to Computational Finance in R*. R package version 1.0/r23.