UNIVERSIDADE FEDERAL DO PARANÁ
FILLIPE RAFAEL BIANEK PIERIN

ANÁLISE DE FUNDOS IMOBILIÁRIOS E ÍNDICES APLICANDO MACHINE LEARNING E MÉTODOS ESTATÍSTICOS

CURITIBA
2021

UNIVERSIDADE FEDERAL DO PARANÁ
FILLIPE RAFAEL BIANEK PIERIN

ANÁLISE DE FUNDOS IMOBILIÁRIOS E ÍNDICES APLICANDO MACHINE LEARNING E MÉTODOS ESTATÍSTICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Matemática Industrial da Universidade Federal do Paraná como requisito à obtenção do título de Bacharel em Matemática Industrial.
Orientador: Prof. Dr. Lucas Garcia Pedroso.
Coorientador: Prof. Dr. Luiz Carlos Matioli.

CURITIBA
2021

Agradecimentos

Primeiramente, agradeço a Deus pela sabedoria e força que me concedeu para tomar as decisões necessárias e para suportar todos os momentos.

Agradeço aos meus pais, por me darem o apoio necessário durante a realização de mais essa etapa da minha vida acadêmica.

Aos colegas de graduação que durante essa jornada estiveram presentes compartilhando conhecimentos e auxiliando mutuamente.

Ao meu orientador, Lucas Garcia Pedroso, e ao meu coorientador, Luiz Carlos Matioli, pela dedicação e paciência que tiveram durante a realização deste trabalho, sendo compreensivos comigo e me ensinando a superar os obstáculos para alcançar os objetivos planejados.

Por último, agradeço aos professores do departamento de Matemática pelos ensinamentos e contribuição para a minha formação acadêmica e profissional.

Resumo

Nesta monografia, apresentamos estudos em fundos imobiliários tratando a parte teórica destes, assim como os tipos existentes e métricas para analisá-los. Realizamos duas análises numéricas, sendo a primeira de previsão de tendência através de *Machine Learning* com a utilização dos modelos Regressão Logística, Floresta Aleatória, Máquina de Vetor Suporte, XGBoost e Rede Neural. E a segunda análise de comparação do retorno e risco dos fundos imobiliários usando o índice IFIX através de testes de hipóteses paramétricos e não paramétricos: Kolmogorov-Smirnov (para verificar suposição de normalidade), T e Wilcoxon (para comparação média), e F e Levene (para comparação da variância). Na análise de tendência, utilizamos dados dos índices Ibovespa e IFIX em duas análises separadas, adicionando variáveis de mercado, alguns índices e cotações, e variáveis obtidas por feature engineering, como coeficiente angular de índices e variação dos valores de fechamento de três meses. Além disso, testamos a retirada de variáveis usando a Análise de Variância (ANOVA) que trouxeram melhoria nos resultados. Os melhores modelos, considerando a medida de avaliação acurácia balanceada, foram Máquina de Vetor Suporte com previsão de um ano à frente com 79.49% usando dados dos índices mais dados de mercado, além de 90% das variáveis mais significativas selecionadas com a ANOVA e XGBoost com previsão de um ano à frente com 75.00% usando as variáveis de mercado, *feature engineering* e 95% das variáveis mais significativas selecionadas com ANOVA, para as análises com os índices IBOV e IFIX, respectivamente. Na análise de comparação, realizamos a comparação do índice IFIX com os índices IBOV, IDIV, SMLL e IMOB, e separamos essa análise em outras considerando meses com IBOV em alta e baixa, e IBOV e IFIX em alta e baixa simultaneamente. Os resultados mostram o índice IFIX comparado com os outros índices da análise possui retorno significantemente igual, e que a volatilidade é significantemente diferente. A exceção ocorre com os índices IDIV e IMOB, que não possuem retorno estatisticamente igual ao índice IFIX quando utilizamos dados quando o índice IBOV estava em período de alta e baixa, respectivamente. Por último, na análise com IBOV e IFIX em baixa, obtemos que o índice IFIX comparado com o índice IBOV apresentam volatilidade igual, o que mostra que estes índices, neste cenário, oscilam de forma semelhante

Palavras-chaves: *Fundos Imobiliários. Índice de Fundos Imobiliários. Ibovespa. Machine Learning. Testes Estatísticos.*

Abstract

In this monograph, we present studies on real estate funds dealing the theoretical part, as well as the existing types and metrics to analyze them. We realize two numerical analyses, the first of trend prediction through *Machine Learning* using the models Logistic Regression, *Random Forest*, Support Vector Machine, XGBoost and Neural Network. And the second of comparison analysis of real estate funds of return and risk of funds using the IFIX index through parametric and non-parametric tests: Kolmogorov-Smirnov (to examine normality assumption), T e Wilcoxon (for average comparison), and F e Levene (for variance comparison). In trend analysis, we are using data from Ibovespa and IFIX indexes in two separate analyses, adding market variables, some indexes and quotation, and variables obtained by feature engineering, such as indexes slope and variation of three month closing values. Furthermore, we test the removal of variables using Analysis of Variance (ANOVA), that brought improvements to the results. The best models, considering the balanced accuracy evaluation metrics, were Support Vector Machine with one year ahead forecast with 79.49% using index data, more market data, beyond to 90% of the most significant variables selected with ANOVA and *XGBoost* with one year ahead forecast with 75.00% using index data, market data and *feature engineering* , beyond to 95% of the most significant variables selected with ANOVA, for the analyzes with the IBOV and IFIX, respectively. In the comparison analyzes, we realize the compared with IFIX index with IBOV, IDIV, SMLL and IMOB indexes, and we separate this analysis into others considering months with IBOV in high and low, and IBOV and IFIX in high and low simultaneously. The results show that the IFIX index compared to the other indexes in the analysis has significantly equal return, and that the volatility was significantly different. The exception occurs with the IDIV and IMOB indexes, which do not have a statistically equal return to the IFIX index when we use data from when the IBOV was in a high and low period, respectively. Finally, in the analysis with IBOV and IFIX indexes in low, we find that IFIX index complied with IBOV index present equal volatility, which show that the indexes, in this scenario, oscillate in a similar way.

Keywords: *Real Estate Funds. Real Estate Funds Index. Ibovespa. Machine Learning. Statistical tests.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

No mercado financeiro há, basicamente, investimentos em renda fixa e variável. Em renda fixa, o investidor poderá saber qual será a rentabilidade do investimento. Porém, em renda variável não sabemos qual rendimento obteremos, pois estes investimentos não possuem um prazo estipulado variando com o tempo e podendo ser impactado pela tendência do mercado e dos investidores. Os investimentos em renda variável costumam ser mais arriscados, no qual o investidor corre o risco de perder parte do valor investido [55].

Antes de investir é indicado que a pessoa, iniciante em investimentos, realize um teste para verificar qual seu perfil de investidor. Esse teste normalmente é realizado quando o investidor cria uma conta em uma corretora, que é o primeiro passo para investir em renda variável e alguns tipos de renda fixa.

O perfil do investidor é dividido em conservador, moderado ou agressivo (arrojado), cujos perfis são definidos levando em consideração a tolerância a riscos [54]. O investidor conservador deseja segurança em seus investimentos, buscando não correr riscos. Para esse perfil de investidor indicamos os investimentos Tesouro Direto, CDB (Certificado de Depósito Bancário), LC (Letra de Câmbio), LCI (Letra de Crédito Imobiliário), LCA (Letra de Crédito de Agronegócio), entre outros. O perfil moderado aceita um pouco de risco, mas de forma controlada, ou seja, o investidor possui segurança em aplicações de renda fixa e começa a investir em renda variável. Por último, o investidor arrojado aceitaria riscos maiores, levando em consideração que as perdas que sofre são momentâneas e que a longo prazo os lucros podem ser maiores, mesmo sem haver garantias.

Dentre os investimentos de renda variável existem os fundos imobiliários, os quais, na literatura da área, são chamados também de FIIs. Os fundos imobiliários são negociados na Bolsa de Valores Brasileira B3 (Brasil, Bolsa, Balcão) [3], antigamente conhecida como BM&FBovespa, que é uma das principais empresas de infraestrutura do mercado financeiro. Para avaliarmos o desempenho dos fundos imobiliário existe o Índice de Fundos Imobiliários (IFIX), que é um índice de fundos relacionados a investimentos no mercado imobiliário, tendo sido inaugurado em 3 de Setembro de 2012 pela B3.

1.1 Justificativa

Os investimentos no setor imobiliário têm crescido nos últimos anos, em razão do valor da taxa Selic estar em baixa (2020) e, também, porque os FIIs apresentam uma volatilidade menor em relação às ações e outros tipos de investimentos [7].

Pela Figura 1.1, percebemos que o índice IFIX apresenta menor volatilidade quando comparado com o índice Ibovespa. Quando existe perda no índice IFIX, cuja curva está em queda, o índice tem uma aparente recuperação da perda um pouco mais rápida. Esses motivos tornam mais atrativos os investimentos nesta modalidade para investidores que não estão habituados.

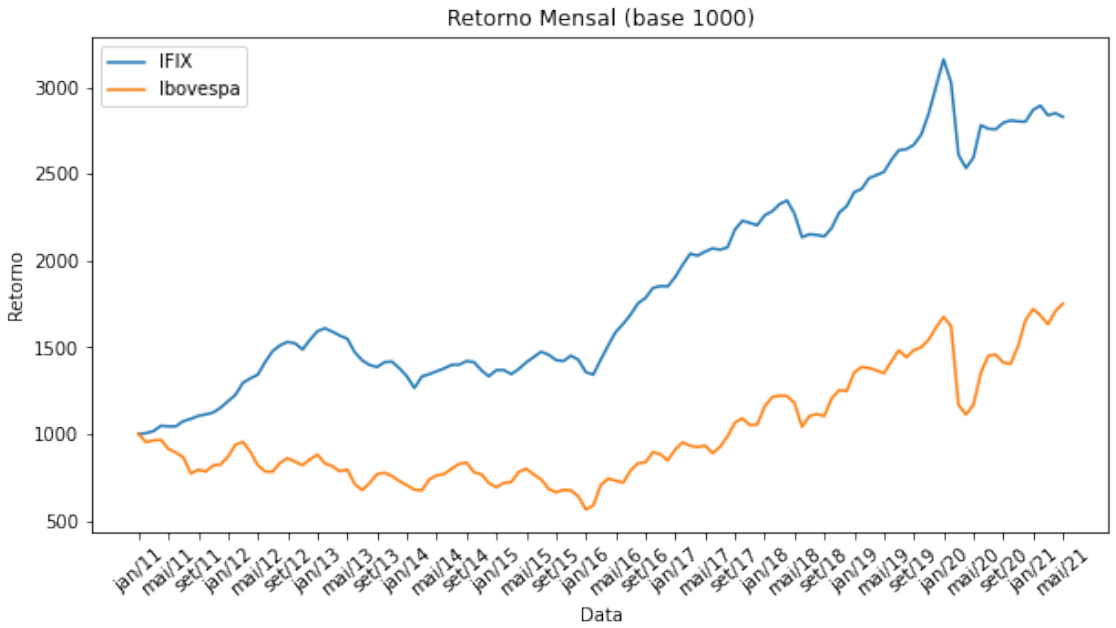

Figura 1.1: Série dos índices Ibovespa e IFIX de Janeiro de 2011 a Maio de 2021 na base 1000.
Fonte: O autor (2021).

A série na base x é utilizado para que a curva de duas séries iniciem com um certo valor em comum [23]. Os valores de uma série na base x é calculada da seguinte forma

$$I_{0,i} = \left(\frac{X_i}{X_0} \cdot x \right), i \in \{1, \ldots, n\}$$

sendo X_i o valor observado da série no período i. Como na Figura 1.1 que temos as séries IBOV e IFIX na base 1000.

A área de Ciência de Dados vem crescendo atualmente, e métodos de Aprendizagem de Máquinas (*Machine Learning*), estão sendo altamente utilizados nas empresas de diversas áreas: saúde, financeira, esportiva, etc. Por isso, ponderamos o uso destas técnicas para analisar os fundos imobiliários a partir do índice IFIX. No Capítulo 6, utilizamos estas técnicas para análises de altas e baixas do índice IFIX, assim como do Ibovespa, o qual é o principal índice de ações do mercado brasileiro.

Além disso, não podemos desconsiderar a Estatística com suas medidas. A Estatística compreende desde a coleta e análise dos dados, até a tomada de decisão. Como no caso da Ciência de Dados, a Estatística é necessária na análise descritiva para concepção das hipóteses que, posteriormente, serão usadas a partir dos dados nos métodos de Aprendizagem de Máquinas. Pela importância verificamos, com base na literatura, que analisar a volatilidade do índice IFIX dentre outros índices usados no mercado seria uma

boa ideia. Para isso, comparamos métricas de índices usando testes de hipóteses, como, em geral é apresentado na literatura da área.

Neste trabalho realizamos análises de fundos imobiliários para uma tomada de decisão sobre a performance do índice IFIX quando comparado com alguns índices e no momento em que são realizadas previsões dos mesmos.

1.2 Objetivos do trabalho

1.2.1 Objetivo Geral

O objetivo geral do trabalho é o estudo de fundos imobiliários, os quais são chamados também, na literatura da área, de FIIs. Buscamos realizar duas análises: uma para prever a tendência de subida ou descida dos índices Ibovespa e IFIX daqui a h meses e uma outra para verificamos se o retorno do índice IFIX é estatisticamente igual a outros índices, além de averiguamos se a volatilidade do IFIX é menor que de outros índices de investimentos. Os índices aqui analisados têm relação aos investimentos em fundos imobiliários e ações. São eles Índice Bovespa (IBOV), Índice de Dividendos (IDIV), Índice *Small Caps* (SMLL) e Índice Imobiliário (IMOB).

1.2.2 Objetivo Específico

1. Estudar os conceitos gerais dos fundos imobiliários:
    - tipos de fundos imobiliários que existem,
    - indicadores usados para analisar os FIIs,
2. Análise de tendência:
    - estudar os métodos de *Machine Learning* (ML), principalmente aprender a utilizar o método de Redes Neurais,
    - estudar as métricas de avaliação usadas nos modelos de ML,
    - aplicar os métodos ML aos dados do índice IFIX, considerando o caso como um problema de classificação,
    - para cada um dos itens anteriores aperfeiçoar o aprendizado de como aplicar na linguagem de programação Python,
3. Análise de comparação de índices:
    - estudar conceitos estatísticos de testes de hipóteses para comparação de média e variância, para sabermos se os retornos são semelhantes entre os índices e para identificamos se a volatilidade de um para outro índice apresenta performance diferente,
    - entender como usar testes paramétricos e não paramétricos na linguagem de programação Python,

- usar os testes de hipóteses para verificarmos se os dados de índices possuem média e variância estatisticamente iguais, isto é, para verificar se os retornos são estatisticamente iguais e se existe variabilidade em torno da média dos dados implicando no índice possuir maior ou menor volatilidade.

1.3 Disposição do Trabalho

A monografia está organizada em em sete capítulos, como apresentamos a seguir. No Capítulo 2, exibimos uma revisão da literatura dos trabalhos existentes que se referem a Aprendizagem de Máquinas usando índices de finanças e trabalhos sobre o uso de testes de hipóteses para análise ou comparação de índices financeiros. No Capítulo 3, apresentamos tipos de fundos imobiliários, que são tijolo e papel, além de detalhar os tipos de fundo por seu objetivo. No Capítulo 4, mostramos os indicadores qualitativos e quantitativos que em geral são utilizados, para analisar e escolher os fundos imobiliários para formar uma carteira. No Capítulo 5, exibimos a metodologia usada nas duas análises com os fundos imobiliários, além da descrição dos dados usados. No Capítulo 6, realizamos a análise de tendência de subida ou descida dos índices Ibovespa e IFIX. No Capítulo 7, efetuamos análises de comparação entre os índices IFIX, IBOV, IDIV, SMLL e IMOB para verificarmos a igualdade de média dos dados, além de averiguamos a hipótese de volatilidade do índice IFIX em comparação aos demais índices. Por último, apresentamos as conclusões.

Algumas definições e termos mais comumente usados na literatura, são apresentadas no Apêndice.

Capítulo 2

Revisão da Literatura

As pessoas, em geral, poupam dinheiro fazendo investimentos hoje para ter mais capital no futuro. Além disso, na velhice os gastos aumentam e é necessário ter mais capital para ter uma boa qualidade de vida. Na área de investimentos existem muitas pesquisas e algumas tentam prever o valor que certa ação terá no futuro para saber qual ação comprar ou vender, também existem pesquisas tentando descobrir a tendência de subida ou descida, porém todas essas pesquisas são apenas previsões do que pode ocorrer, visto que os investimentos, principalmente em renda variável como fundos imobiliários e ações, são muito voláteis e, dependendo do período, crises e impactos na economia podem mudar o cenário e a tendência destes investimentos. Desta forma, a volatilidade é uma variável importante a ser analisada na literatura.

No trabalho [21], Finkler apresenta um estudo usando técnicas de *Machine Learning* para realizar a previsão dos movimentos do índice IBOV, ou seja, realiza a previsão da tendência de subida ou descida no horizonte de h meses à frente, com $h \in \{1, 3, 6, 12\}$, e para isso utiliza dados históricos mensais do índice IBOV do período de Janeiro de 2002 à dezembro de 2016. O estudo foi realizado como um problema de classificação com duas classes: -1 quando o índice IBOV caiu em h meses à frente e 1 quando ele subiu. As técnicas de *Machine Learning* empregadas foram Regressão Linear, Regressão Logística, Máquina de Vetores Suporte com margens flexíveis (C-SVM) e Rede Neurais Artificiais, além de uma combinação entre os modelos. O melhor resultado obteve a taxa de acerto de $72,7\%$ usando o modelo C-SVM para predição num horizonte de 6 meses. Porém, usando uma combinação híbrida dos modelos Regressão Linear, Regressão Logística e C-SVM, a autora obtém uma taxa de acerto de $78,8\%$ (melhora de $6,1\%$). Por fim, na dissertação a autora realiza uma comparação da técnica de *Machine Learning* com a estratégia do tipo *buy and hold* usando investimento do fundo de índices de ações BOVA11 (carteira de ativos com base no índice Ibovespa) com atualização a cada semestre, cujo resultado mostra que o uso de técnicas de *Machine Learning* resulta em retornos mais significativos que a estratégia *buy and hold*.

Santos [60] em sua dissertação de mestrado exibe uma análise para prever a tendência de três ativos negociados na bolsa de valores brasileira (B3): Vale (VALE3), Petrobras (PETRA4) e Itaú Unibanco (ITUB4). Para essa finalidade o autor aplicou os seguintes algoritmos de *Machine Learning*: Floresta Aleatória (*Random Forest* - RF), Redes Neurais Artificiais (RNA) e Máquina de Vetores Suporte (*Support Vector Machine* - SVM), e também buscou investigar a eficiência dos modelos para previsão de 30, 60 e 90 dias úteis.

Além disso, Santos cita que foram utilizados cinco indicadores técnicos, que são: índice de força relativa, oscilador estocástico, Williams %R, convergência e divergência da média móvel e taxa de mudança de preços. Os dados utilizados pelo autor consideram o período de 11 de Fevereiro de 2015 a 10 de Fevereiro de 2020. Segundo o autor "a volatilidade típica do mercado de ações torna a tarefa da previsão mais difícil", o que mostra que esse tipo de previsão é complicada e trabalhosa. O resultado obtido por Santos foi que a acurácia compreende entre 71% e 97% quando aplicado a separação dos dados em treinamento e teste de forma aleatória, já na ocasião onde a separação dos dados ocorre de forma temporal, a acurácia obtida varia entre 18% e 76%. Dentre estes resultados, os melhores com divisão dos dados aleatória para cada cenário foram

- 96% para o algoritmo RF, ativo VALE3 e previsão 30 dias,
- 97% para o algoritmo SVM com kernel RBF e previsão 90 dias,
- 97% para o algoritmo SVM com kernel RBF, ativo PETR4 e previsão 90 dias,

e para o estudo com divisão de dados temporal foram

- 68% para o algoritmo SVM com kernel GHI, ativo VALE3 e previsão 90 dias,
- 77% para o algoritmo SVM com kernel RBF e previsão 90 dias,
- 64% para o algoritmo RNA, ativo PETR4 e previsão 60 dias,

Por último, o autor conclui que apesar de os resultados serem melhores utilizando divisão dos dados de forma aleatória, estes resultados não se aplicam a vida real porque os dados são temporais.

Serra e Morais [69] analisam duas classes de ativos, FIIs e ações, comparando o retorno do índice IFIX com os retornos dos índices IBOV, IDIV, SMLL e IMOB, empregando testes de hipóteses. Segundo os autores, as comparações têm em vista entender características dos FIIs, auxiliando investidores, gestores e pessoas interessadas em realizar investimentos em FIIs e desejam entender as aplicações na área imobiliária. Os testes utilizados foram de igualdade de média e igualdade de variância, tanto paramétrico quanto não paramétrico, conforme a configuração dos dados. Os dados de forma mensais utilizados compreendem o período de 30 de Dezembro de 2010 a 31 de Agosto de 2017 e a análise foi separada em três partes cujos dados mensais utilizados foram: todos os dados (análise 1), IBOV em alta e IBOV em baixa (análise 2), e por último IFIX e IBOV em alta e IFIX e IBOV em baixa (análise 3). Os autores concluíram na análise 1 que o IFIX tem retornos estatisticamente iguais aos demais índices confrontados, mas que não possui a característica teórica esperada de possuírem risco semelhante. Na análise 2, o resultado obtido foi que o IFIX mostrou ter menor risco em comparações com os outros índices considerados, além de que o IFIX subiu menos em tempo de alta e caiu menos em tempo de baixa, em comparação com os demais índices. Esse resultado diferiu aos da análise com todos os dados. Com a análise 3, o resultado obtido foi que o IFIX subiu menos nos meses de alta e teve uma queda menor que o IBOV nos meses de baixa, tendo desempenho de um ativo de menor risco. No geral, os autores concluíram que o índice IFIX tem um comportamento diferente das ações, tendo um risco inferior confrontado aos outros índices e o retorno diferem quando as análises são detalhadas, o que mostra que a volatilidade reflete no retorno.

No artigo [11], os autores executaram um estudo com o objetivo de explorar o desempenho do Índice de Desenvolvimentos em comparação a outros cinco grupos de índices presentes na B3. Os seis grupos de índices considerados são: de Sustentabilidade, Amplos, Setoriais, de Governança, de Segmentos e Outros, sendo que há vinte e quatro índices no total dos grupos. Para essa comparação foram utilizados testes estatísticos conhecidos como testes de hipóteses. Os testes de hipóteses utilizados foram de igualdade de média não paramétricos Wilcoxon e Mann-Whitney, quando não havia normalidade nos dados, e o teste paramétrico t de student, quando o pressuposto de normalidade era aceito. Para a análise da variância utilizaram o teste paramétrico ANOVA e o teste não paramétrico Levene para cada situação. Também foram utilizados os testes Welch, Brown-Forsythe e Kruskal Wallis quando os dados tiveram resultado de heterocedasticidade. Além destes testes, foram aplicados os testes Kolmogorov-Smirnov e Shapiro-Wilk para verificar a suposição de normalidade. Dos resultados, os autores, obtiveram que os índices possuem retorno médio diário estatisticamente iguais, mas com as medianas diferenciando com o emprego do teste Kruskal Wallis, para dados do período analisado. Além disso, que o uso dos testes Wilcoxon e Mann-Whitney resultaram não haver diferença estatística entre o grupo de Sustentabilidade e os outros grupos. Em consequência desse resultado, os autores concluem que não existe diferenças relevantes em realizar investimentos em um ou outro índice, para o período analisado.

Capítulo 3

Tipos de Fundos Imobiliários

Os FIIs são divididos em diferentes classes disponíveis para o investidor. Porém, essa classificação pode se dar de duas formas, conforme o tipo ou pelo objetivo para obter renda [10].

Pelo tipo, existem os fundos de tijolo e de papel.

1. **Tijolo**: são fundos que investem em imóveis físicos. Os investidores ganham com a renda de aluguéis ou venda desses imóveis;

2. **Papel**: são fundos que investem em cotas de outros fundos, em ativos financeiros ou ambos. Estes fundos são compostos por ativos intocáveis como depósitos bancários, obrigações e ações, em que os investidores ganham com o pagamento de juros.

Em relação à classificação por objetivo, temos mais de dez opções de investimento em fundos imobiliários [25], que vão de investimentos em agências bancárias, shoppings, lajes corporativas a investimentos em Certificado de Recebíveis Imobiliários (CRI) e Letra de Crédito Imobiliário (LCI). A seguir abordaremos cada uma das possibilidades.

- **Fundos de Renda**: são fundos que constroem ou compram imóveis para alugar e os cotistas recebem o valor relativo aos aluguéis. Esses fundos de renda costumam investir em:

  - **Shoppings e Varejo**: shoppings e lojas de varejo, sendo bastante diversificados, visto que há várias lojas em um shopping. Porém, esses investimentos dependem da lucratividade das lojas, podendo gerar perda de rendimentos por parte dos fundos nos períodos de baixa;
  - **Lajes Corporativas**: lojas de alto padrão alugadas para grandes empresas, que possuem na maioria das vezes aluguel fixo e corrigido pela inflação. Esses empreendimentos podem ter um ou mais inquilinos;
  - **Galpões Industriais**: são galpões alugados para, em geral, um único inquilino. Os FIIs podem possuir um ou mais ativos de galpões;
  - **Imóveis Residenciais**: imóveis com inquilinos, que podem ser mais diversificados. Os investimentos de FIIs com imóveis residenciais costumam ser mais arriscados, pois depende dos aluguéis dos inquilinos;

  - **Agências Bancárias**: são investimentos em imóveis alugados para bancos. Os fundos investidos nestes tipos de imóveis, normalmente, possuem um único proprietário (banco). Os aluguéis são fixos e ajustados pela inflação;
  - **Escolas, Universidades e Hospitais**: fundos com um ou mais imóveis, em que os rendimentos são condicionados à receita do locatário. Neste caso, o locatário cede o imóvel para criação de área hospitalar ou de estudo;
  - **Hotéis e flats**: investem em flats e hotéis, e os rendimentos estão vinculados aos rendimentos das unidades, que podem afetar conforme o momento do setor;
  - **Híbridos (tijolos e papéis)**: fundos que investem tanto em fundos do tipo de tijolo quanto de papel. No portfólio deste pode haver CRI, LCI, fundos de fundos e os demais tipos de fundos de renda citados acima. Desta forma, a diversificação é obtida sem muitos custos e com menos riscos;

- **Fundos de Desenvolvimento**: são fundos que investem na construção de imóveis para venda. Esses fundos lucram com a venda desses imóveis, sendo um ativo de alto risco;

- **Fundos de Compra e Venda**: são fundos que lucram com a venda e compra de imóveis, cujos rendimentos são variáveis;

- **Fundos Certificado de Recebíveis Imobiliários (CRI)**: são fundos que possuem títulos de renda fixa. Esses títulos equivalem a uma promissória de recebimento de um certo valor no futuro [24]. Os CRI são de prazo longo, correspondendo geralmente a um período de quatro a dez anos, podendo chegar a quinze anos;

- **Fundos de Letras de Crédito Imobiliários (LCI)**: são fundos com aplicação em renda fixa gerada pelos bancos. Essas aplicações são isentas de imposto de renda e, em geral, rendem mais que a poupança. Um tipo de aplicação parecida é a Letra do Crédito Agropecuário (LCA), que diferentemente do LCI possui investimentos em empréstimos ofertados por produtores rurais e cooperativas [32];

- **Fundos de Fundos**: são fundos que possuem rendimentos a partir dos investimentos em outros FIIs. Desta forma, são fundos bastante diversificados. Estes fundos são ideais para os investidores que querem não ter a preocupação de escolher os ativos para montar sua carteira.

Capítulo 4

Principais Conceitos Envolvidos na Análise de Fundos Imobiliários

Neste capítulo, iremos apresentar alguns indicadores que são comumente utilizados na análise de fundos imobiliários, tanto indicadores da análise qualitativa quanto da quantitativa. Para exemplificação, usaremos dados do fundo imobiliário CSHG Real Estate, cujo código é HGRE11 com relação aos dados de Julho de 2020 a Junho de 2021 [14]. Este fundo tem como classificação tijolo e compreende investimentos em lajes corporativas.

No Anexo D, temos a relação dos 20 imóveis (107 unidades no total) que compõe o fundo HGRE11, exibindo os principais indicadores e a classificação de cada imóvel, na época em que realizamos o estudo (ano de 2021).

4.1 Análise Qualitativa

Segundo Tim Smith [72], a análise qualitativa se fundamenta em explorar os princípios de fundos ou empresas a partir de referências não numéricas. Deste modo, a análise qualitativa pode ser intangível ou inexata pelas informações analisadas serem mais subjetivas. Levando em consideração os fundos de investimentos, a partir da análise qualitativa constatamos as possíveis ameaças e oportunidades e verificamos os atributos dos locatários dos imóveis e o risco de inadimplência entre eles e os contratos dos fundos [18].

4.1.1 Escolha do segmento da aplicação

A primeira coisa a realizar quando desejamos investir em FIIs é a escolha do segmento ou tipo de fundo entre os existentes, que explicitamos na seção anterior, para conseguirmos uma maior diversificação. A diversificação é necessária, pois alguns setores podem estar em baixa devido às condições de mercado e outros em alta, e na média o investidor corre menos risco de perdas.

4.1.2 Examinando a condição de um ativo

Para examinarmos a condição de um ativo precisamos verificar se o(s) imóvel(is) presente(s) no fundo imobiliário em que estamos investindo está(ão) localizado(s) em boa região; analisar os valores dos imóveis da(s) região(ões) do(s) imóvel(is) presente(s) [76]

e apurar se a taxa de vacância não está alta nem baixa, isto é, verificar a quantidade de imóveis e inquilinos para saber se a renda não está sendo ou será comprometida.

4.1.3 Aferindo se um fundo imobiliário é monoativo ou multiativo

Os fundos imobiliários monoativos possuem referência a fundos aplicados em imóvel único, como shopping centers [50], ou seja, a renda provém de um único ativo. Todavia, se tivermos mais de um imóvel no fundo imobiliário estaremos com um fundo multiativo, tendo diversificação dos ativos.

Na prática podemos ter fundos imobiliários em que os imóveis possuem um ou mais inquilinos. Os imóveis com poucos inquilinos podem acarretar em imóveis com parte vaga, podendo gerar perda de renda.

4.1.4 Averiguando a localização do imóvel

A averiguação da localização dos imóveis, presentes no fundo, é necessária para sabermos se o valor aplicado dos aluguéis e compras dos mesmos estão adequados para a região. Ou seja, se não existe cobrança indevida nos valores abaixo ou acima dos cobrados na região, prejudicando ou beneficiando o(s) proprietário(s) do(s) imóvel(is).

Também é pertinente analisarmos se a região onde estão os imóveis do fundo possuem potencial de valorização ou desvalorização, para verificarmos possível aumento ou decréscimo da renda desse fundo no futuro.

4.1.5 Analisando a gestão e o administrador

Outro fator que precisamos verificar é a qualidade da gestão e do administrador, para isso devemos analisar como foi a administração nos últimos anos do fundo imobiliário em questão, com a finalidade de saber se houveram prejuízos ou benefícios para os investidores deste fundo.

É fundamental verificarmos o histórico do gestor do fundo, para averiguarmos se no passado teve boas administrações com relação aos investimentos monetários. Também é ideal analisarmos o histórico do administrador, para sabermos se os serviços prestados pela empresa ou corretora que cuida do fundo foram bem gerenciados.

Após a análise qualitativa, é necessário também realizarmos a análise quantitativa apresentada na próxima seção, para uma tomada de decisão mais completa e abrangente. Além disso, nenhum dos dois tipos de análises são independentes, mas complementares.

4.2 Análise Quantitativa

Na análise quantitativa consideramos alguns indicadores para escolher os melhores fundos imobiliários para se investir [8]. Na prática estes indicadores são mais usados porque a comparação numérica é mais confiável que a análise qualitativa e não acaba levando em consideração critérios ou ideias subjetivas. Neste caso abordaremos os dez principais indicadores exibindo como funcionam na prática.

4.2.1 Vacância

A vacância é um índice usado para averiguarmos a porcentagem de área do(s) imóvel(is) presente(s) no fundo que não está alugada [74]. Esse indicador é dividido em vacância física e financeira. A vacância física compreende a porcentagem de área não alugada de um negócio, diferente de taxa de ocupação. Já a vacância financeira se refere à quantidade de fluxo de caixa gerado pelo fundo. Por exemplo, no caso do fundo HGRE11, no mês de Junho a vacância física foi de 22.34% e a vacância financeira de 24.48%, para meses anteriores ver Figura 4.1.

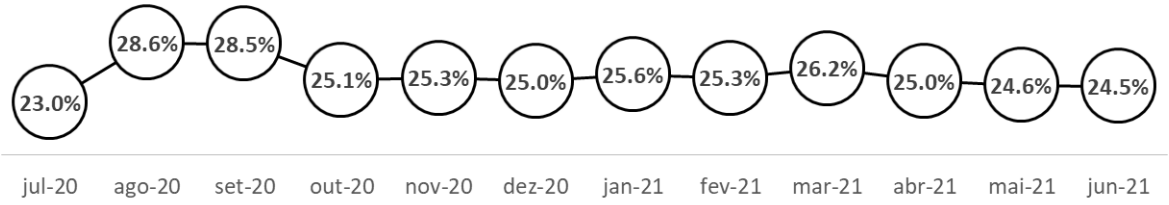

Figura 4.1: Dados do fundo imobiliário HGRE11 com relação a medida de vacância financeira.
Fonte: Adaptado de Credit Suisse Hedging-Griffo (2021).

4.2.2 *Dividend Yield* (DY)

O *Dividend Yield* é uma medida usada para sabermos se um fundo está repassando os lucros aos cotistas de forma adequada, ou seja, o dividendo anual pago (por cota) pelo valor unitário do ativo. O dividendo é a parte dos lucros repassada aos cotistas, que no caso dos FIIs representa 95.0% do lucro. Além disso, esse rendimento nos fundos imobiliários ocorre mensalmente. Por exemplo, em relação ao fundo HGRE11, o *dividend yield* é de 12.47% (R$1.38 x 12 / R$132.85) anual em Junho de 2021 [14]. Os rendimentos correspondentes a locação, valores imobiliários e ganhos de capital bruto (inclui venda de imóveis, fundos e outros ativos imobiliários) do fundo HGRE11 nos doze meses anteriores foi de R$140054249.00 e em Junho foi de R$21509959.00 [14].

4.2.3 Número de Cotas

Uma cota de FII representa a fração ideal de seu patrimônio. Quando compramos uma cota de um FII tornamo-nos sócios de parte de empreendimentos imobiliários como shoppings, galpões comerciais, lojas, etc. Em todos os meses do período de Julho de 2020 a Junho de 2021, o número de cotas do fundo HGRE11 foi de 11817767 [14], sendo que o preço da cota em 30 de Junho de 2021 estava em R$132.85.

4.2.4 Valor Patrimonial do Fundo

O valor patrimonial (VP) é a representatividade de forma quantitativa do valor efetivo de algum patrimônio. O indicador está relacionado ao conceito de patrimônio líquido, que

é o resultado da diferença dos deveres e obrigações, e dos bens e direitos [53]. O valor patrimonial é dado pela expressão

$$\text{VP} = \frac{\text{PL}}{\text{número de ações}}$$

em que VP é o valor patrimonial e PL é o patrimônio líquido.

Considerando o fundo HGRE11, temos que em Junho de 2021, que o patrimônio líquido era de R$1993900000.00 [14] e, como visto anteriormente, no mesmo mês o número de cotas desse fundo era de 11817767.00. Isto implica no valor patrimonial de R$168.72.

4.2.5 Preço sobre Valor Patrimonial da Cota

O preço sobre valor patrimonial (P/VP) da cota indica a depreciação ou aumento do valor nominal do ativo que está sendo negociado [76].

O preço da cota do fundo HGRE11 era de R$ 132.85 em 30 de Junho de 2021 e o valor do índice VP que mostramos anteriormente é de R$168.72. Desta forma, obtemos que o valor do P/VP da cota é de R$132.85 / R$168.72 = 0.79. Logo, temos que o mercado está precificando em -21% o valor desse fundo, isto significa que existe uma depreciação do valor do ativo.

4.2.6 Cap Rate

O *cap rate* é um índice que reflete a quantidade de renda angariada por um imóvel por ano [75]. Ou seja,

$$\textit{cap rate} = \frac{\text{total arrecadado}}{\text{valor avaliado do imóvel}},$$

podendo ser calculado para diferentes períodos: mensal, trimestral, anual, etc.

Por exemplo, tendo em vista um imóvel avaliado em R$251200.00 ($50m^2$) com aluguel de R$1200.00 mensais, obtemos um *cap rate* anual de 5.73% $\left(\frac{R\$1200.00 \cdot 12}{R\$251200.00} \cdot 100\right)$. O que significa que o investidor terá um retorno de 5.73% com este imóvel.

Porém, devemos tomar cuidado com esse índice, porque este pode ser afetado pela vacância e regulamentação. Isso pode ocorrer, por exemplo, quando alguns dos imóveis do fundo têm poucos inquilinos, ou ainda quando não estão em situação regular nos órgãos públicos, impedindo a manutenção da licença de funcionamento.

4.2.7 Área Bruta Locável (ABL)

Área Bruta Locável (ABL) é uma medida que diz respeito à quantidade de espaço em metros quadrados $\left(m^2\right)$ brutos disponíveis para aluguel. Em geral, o terreno não é considerado para computar o valor da ABL, somente o que é considerado em cima do local, isto é, a área interna do imóvel. Essa métrica é usada nacionalmente para comercialização de propriedades como lojas, galpões logísticos e de varejo, shoppings, etc.

No Brasil, podemos ver em relatórios a ABL ser anunciada a partir das áreas privativas. As áreas privativas são locais dos imóveis onde somente os locatários podem utilizar. No exterior há fundos que consideram o conceito *Building Owners and Managers Association* (BOMA), que compreende a soma da área privativa com o espaço disponível para aluguel [9].

Nas Figuras 4.2 e 4.3, temos os gráficos dos valores da ABL de cada imóvel e ABL por região dos imóveis presentes no fundo imobiliário HGRE11, dados de Junho de 2021 [14]. O edifício empresarial Dom Pedro é o que possui a maior ABL (25543 m^2) entre os imóveis e o edifício Brasilinterpart é o que possui o menor ABL (887 m^2) do fundo. Em relação à região que contém imóveis do fundo, a região Santo Amaro / Chácara Santo Antônio / Morumbi / Chucri Zaidan (SCMC) têm maior ABL (30588 m^2), porque contém mais imóveis que as demais regiões e a região do Rio de Janeiro é a que tem menor ABL (4027 m^2) apesar de ter dois imóveis no fundo, sendo que outras regiões como Curitiba tem apenas um imóvel.

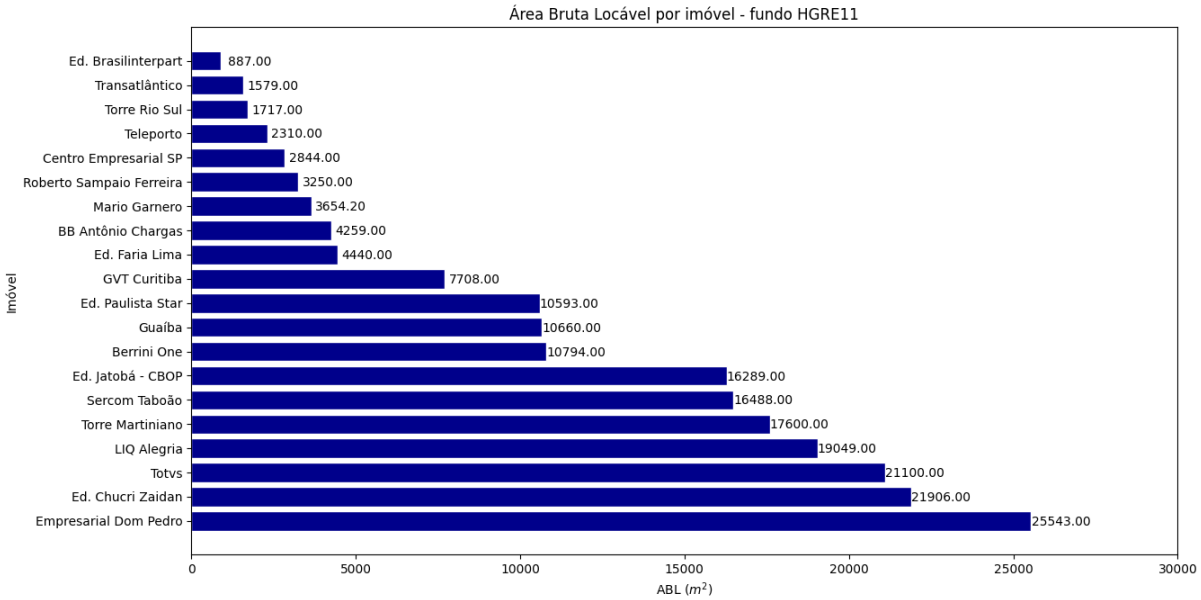

Figura 4.2: ABL dos imóveis que compõem o fundo imobiliário HGRE11.
Fonte: O autor (2021).

4.2.8 Valor do aluguel por m^2

O Valor do aluguel por m^2 é um indicador que utilizamos para sabermos se o valor do aluguel aplicado no(s) imóvel(is) do fundo estão adequados aos valores de outros imóveis da mesma região [76]. Desta forma, o proprietário pode colocar um aluguel mais baixo ou alto, caso verifique que o aluguel está caro ou barato, dependendo da taxa de vacância. A vacância física é inversamente proporcional ao valor do m^2 cobrado. Assim, caso os valores cobrados por m^2 na região sejam discrepantes, melhor suspeitar e averiguar se estão corretos [76].

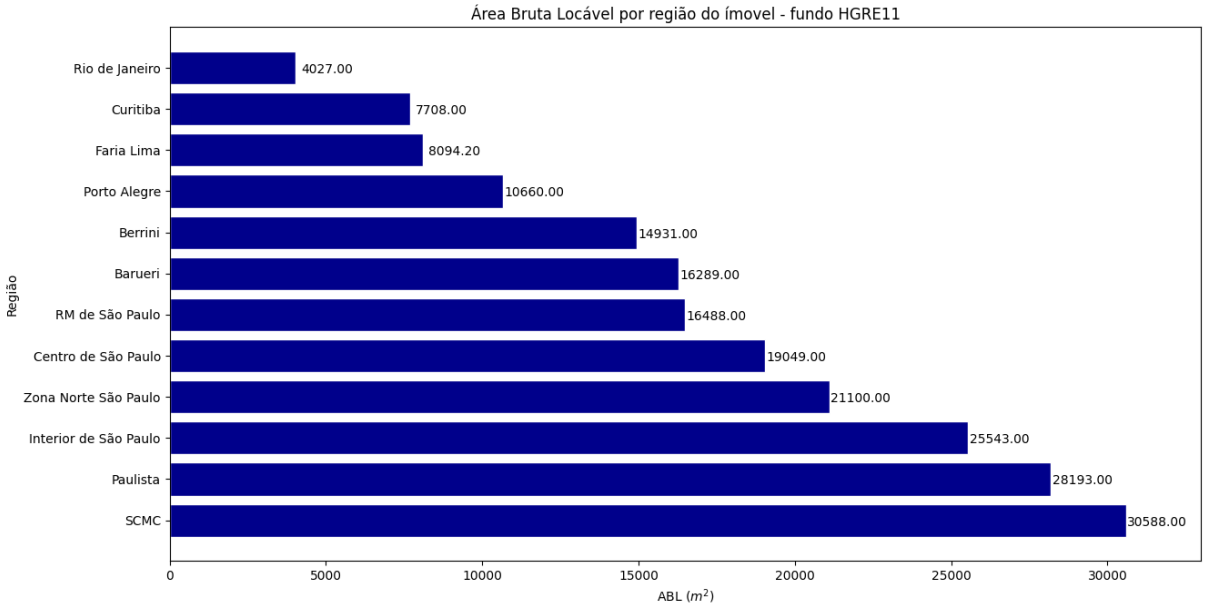

Figura 4.3: ABL por região dos imóveis que compõem o fundo imobiliário HGRE11. Fonte: O autor (2021).

4.2.9 Valor por m^2 do imóvel em relação à cotação de mercado

O valor por m^2 do imóvel em relação à cotação de mercado é um índice normalmente empregado para verificarmos as oportunidades nos valores das cotações negociadas na bolsa de valores, ou averiguarmos se existe aumento no preço de imóveis que estão no fundo. Por exemplo, considerando que um imóvel é avaliado em R$251200.00 ($50m^2$), porém pela cotação na bolsa de valores sairá por R$230150.00 ($50m^2$). Desta forma, verificamos que a negociação na bolsa está barata. Agora suponha que este imóvel possua ABL de 5000 m^2, consequentemente o valor do m^2 dividido por valor patrimonial será de R$5024.00 (R$25.12 milhões / 5000) e baseado na bolsa será R$4603.00 (R$23.02 milhões / 5000). Portanto, cada m^2 do imóvel custará ao investidor R$4603.00 [76].

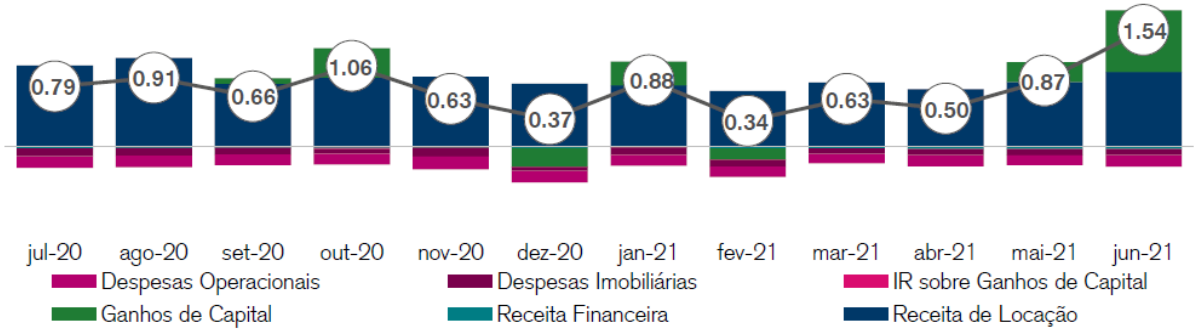

Figura 4.4: Dados do fundo imobiliário HGRE11 com relação a despesas e ganhos. Fonte: Credit Suisse Hedging-Griffo (2021).

4.2.10 Taxas de administração e performance

Nos fundos imobiliários são aplicadas as taxas de administração e performance. A taxa de administração é a taxa obtida pelos gestores e administradores em relação aos serviços de administração dos fundos, nos fundos imobiliários essa taxa é de no máximo 1.0% ao ano. Já a taxa de performance é o valor recolhido quando o fundo atinge um índice estabelecido de antemão na documentação do mesmo [16].

Por exemplo, em relação ao fundo HGRE11 há uma taxa de administração de 1.0% a.a. [14] sobre o valor de mercado do fundo. Para mais valores desse fundo a respeito de despesas imobiliárias, despesas operacionais, imposto de renda (IR), ganhos entre outros, ver Figura 4.4.

Capítulo 5

Materiais e Métodos

5.1 Descrição dos Dados

Nas duas análises propostas neste trabalho, manipulamos bases de dados referentes à índices financeiros e medidas do mercado financeiro. Estas bases de dados são compostas por dados mensais de preços de fechamento dos índices:

- Índice de Fundos de Investimentos Imobiliários (IFIX),
- Índice Ibovespa (IBOV),
- Índice de Dividendos (IDIV),
- Índice *Small Cap* (SMLL),
- Índice Imobiliário (IMOB),
- Cotação do dólar,
- Cotação do ouro,
- Índice Nacional de Preços ao Consumidor Amplo (IPCA).

Nestas bases de dados consideramos os seguintes períodos:

- 01 de Janeiro de 2001 a 31 de Maio de 2021 na análise de tendência usando IBOV, cotação do dólar, cotação do ouro e IPCA,
- 01 de Janeiro de 2011 a 31 de Maio de 2021 na análise de tendência usando o IFIX, IBOV e IPCA,
- 01 de Janeiro de 2011 a 31 de Maio de 2021 na análise de comparação de índices usando IFIX, IBOV, IDIV, SMLL e IMOB.

Optamos por utilizar esse período para contemplar todo o desempenho do índice IFIX desde sua origem em Dezembro de 2010.

Os dados dos índices IFIX e IBOV e da cotação do dólar e do ouro obtivemos do site Yahoo Finance [79]; os dados dos índices IDIV, SMLL e IMOB retiramos do site da

bolsa de valores do Brasil, B3 [3]; por último os dados do índice IPCA puxamos do site do Instituto Brasileiro de Geografia e Estatística (IBGE) [30].

Para a análise destes dados fizemos a utilização da linguagem de programação Python 3.8.3 [58] no ambiente Google Colab. Os pacotes utilizados da programação Python foram os seguintes:

- **SciPy**: para realização dos testes de hipótese,
- **Scikit-learn**: para aplicação dos modelos de *Machine Learning*,
- **Xgboost**: para aplicação do modelo *XGBoost*,
- **Tensorflow** / **Keras**: para a aplicação do modelo de Redes Neurais,
- **Matplotlib** e **Seaborn**: para manipulação e plotagem dos gráficos,
- **Numpy** e **Pandas**: para manuseio de tabelas / *data frames* dos dados utilizados durantes as análises.

Durante a análise de tendência utilizamos os dados brutos de fechamento, que somente normalizamos antes de aplicar os modelos de ML. Na análise de comparação de índices utilizamos o log-retorno entre os instantes $t-1$ e t [46]. Considere P_t o preço de fechamento do índice no instante t. O retorno entre os instantes $t-1$ e t é dado por

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 \quad \Longrightarrow \quad 1 + R_t = \frac{P_t}{P_{t-1}},$$

como os instantes t e Δt são muito próximos, então $R_t << 1$. Sendo assim, para esses casos $ln(1+R_t) \approx R_t$, doravante obtemos que o log-retorno é obtido da seguinte forma

$$R_t \approx ln(1+R_t) = ln\left(\frac{P_t}{P_{t-1}}\right) = ln(P_t) - ln(P_{t-1}).$$

5.2 Índices de Investimentos

Os índices de investimentos são métricas que transmitem as informações de performance de investimentos ou organização, e normalmente, são usados para a tomada de decisões, ou seja, são usados como referência para verificar a rentabilidade e permitir comparações com os resultados do mercado. Neste trabalho, analisamos e comparamos os índices de investimentos IFIX, IBOV, IDIV, SMLL e IMOB.

5.2.1 Índice de Fundos de Investimentos Imobiliários

O Índice de Fundos de Investimentos Imobiliários (IFIX) é o índice que estima a performance média do portfólio teórico dos fundos imobiliários listados na B3 [19]. A liquidez e o valor de mercado são parâmetros usados pela B3 como critério para escolha dos ativos que compõem a carteira teórica. Este índice serve como parâmetro de desempenho e os critérios para seleção dos fundos imobiliários para entrar no IFIX são os seguintes: nas últimas três carteiras ter pelo menos 60% de presença em pregões; as cotas do fundo não

podem ser consideradas como *penny stocks*, ou seja, com cotações abaixo de R$1.00; a oferta pública do fundo precisa ser efetuada antes do balanceamento e suas cotas precisam estar entre os ativos elegíveis, que representam 99% do total dos indicadores que constituem o índice, em ordem decrescente do Índice de Negociabilidade [19]. O IFIX tem em sua composição 103 fundos imobiliários, em 30 de Outubro de 2021, sendo que a carteia teórica é atualizada quatro vezes no ano.

5.2.2 Índice Bovespa

O Índice Bovespa (IBOV) é um indicador que mede o comportamento das principais ações que são negociadas na B3. Esse índice é uma carteira teórica de ações, composta pelas empresas com maior volume financeiro comercializado na B3 [17]. Os critérios para escolha das ações são: empresas que possuem constância, tendo presença em 95% dos pregões do ano anterior; empresas com volume financeiro de pelo menos 0.1% do volume negociado no período; não ser empresa em recuperação judicial; e as empresas não podem ser classificadas como *penny stocks* [17]. A composição da carteira do índice é revisada a cada quatro meses e contém 92 ativos de empresas, em 30 de Outubro de 2021.

5.2.3 Índice de Dividendos

O Índice de Dividendos (IDIV) é um indicador que representa a performance média dos ativos da B3 que pagam de melhor forma seus investidores com dividendos [51]. Desta forma, é basicamente uma carteira de dividendos teórica dos ativos, sendo que a cotação espelha o preço dos ativos que compõe o índice, mostrando o impacto dos dividendos. Segundo dados do site da B3, a carteira de dividendos IDIV é composta apenas por ativos (ações e units) de companhias listadas na B3 num total de 46, em 30 de Outubro de 2021. A composição dessa carteira é atualizada a cada trimestre.

5.2.4 Índice *Small Cap*

O Índice *Small Cap* (SMLL) é uma carteira de ações com papéis de empresas que possuem uma menor capitalização. Este índice permite verificar o comportamento de empresas *small caps*, sendo útil para analisar se as ações estão valorizando ou desvalorizando [22]. O SMLL é composto por 128 ativos, em 30 de Outubro de 2021. Os critérios para escolha da ação na carteira teórica são que o ativo da empresa esteja entre os 99% com mais comercialização na bolsa de valores e fora da lista que representa 85% do valor de mercado das empresas relacionadas na B3; ter presença em 95% dos pregões dos três meses anteriores; e que as cotas não sejam *penny stocks* [36].

5.2.5 Índice Imobiliário

O Índice Imobiliário (IMOB) é uma carteira teórica composta pelos ativos (ações e units), utilizada para medir a performance média dos ativos que possuem maior negociação e possuem representatividade no que diz respeito à atividade imobiliária e são mais negociadas [52]. O IMOB é composto por 26 ativos, em 30 de Outubro de 2021, com atualização trimestral. Para o ativo estar na carteira teórica precisa seguir os seguintes critérios: não se encontrar em situação de recuperação tanto judicial quanto extrajudicial; não ser

classificada como uma *penny stock*; ter no mínimo 95% de negociação nos últimos três pregões da bolsa estando com oferta pública durante esse período; não ter intervenção; o ativo não ter trabalho em regime especial de administração temporária; e não encontrar em situação de listagem especial [52].

Para mais detalhes sobre os índices e suas composições, consultar o site da B3.

5.3 Modelos de *Machine Learning*

A metodologia que utilizamos, durante a análise de tendência, compreende modelos ou métodos de Aprendizagem de Máquinas *Machine Learning* (ML) para averiguarmos a tendência de subida ou descida do índice IFIX t meses à frente. No decorrer da aplicação dos métodos ML empregamos a técnica Validação Cruzada ou Cross Validation (CV) para fazermos a avaliação de desempenho dos modelos. Os modelos de ML usados foram Regressão Logística (RL), Floresta Aleatória (RF), *XGBoost* (XGB), Máquina de Vetores Suporte (SVM) e Rede Neural (RN).

Nos modelos de Aprendizagem de Máquinas, seguimos alguns passos para a realização de classificação ou regressão. Na classificação a variável resposta pertence a um conjunto finito de valores, ou seja, $y_i \in \{0, 1\}$, $i \in 1, 2, \ldots, n$. E na regressão a variável resposta pertence a um conjunto infinito de valores, isto é, $y_i \in \mathbb{R}$, $i \in 1, 2, \ldots, n$. Esses passos englobam a divisão dos dados em treinamento e teste, a aplicação dos dados de treinamento nos modelos para a escolha dos melhores parâmetros por meio de uma medida de avaliação, a predição de novos valores nos dados de testes e a avaliação do possível modelo final usando uma medida de avaliação. Na Figura 5.1 apresentamos o fluxograma com cada um desses passos.

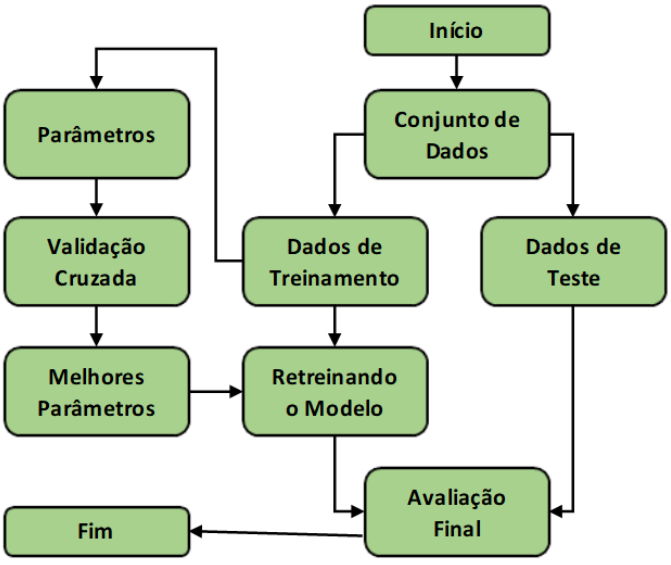

Figura 5.1: Fluxograma do passo a passo da modelagem usando Aprendizagem de Máquinas.
Fonte: Adaptado de [63].

Primeiramente em ML realizamos uma análise descritiva, para identificar possíveis valores nos dados que possam ser discrepantes, normalizamos os dados caso seja necessário,

e selecionamos algumas das variáveis mais importantes em alguns casos. Em seguida, com os dados arrumados, dividimos os dados em duas partes, sendo uma parte para treinamento e outra parte para teste ou validação do modelo utilizado. Essa partição pode se dar em qualquer partição de $x\%$ de dados para treinamento e $100\% - x\%$ para teste, considerando a parte de treinamento sempre maior. Na prática, os valores mais comuns de divisão a se usar é 70% ou 80% para treinamento e o restante para teste. Neste trabalho usaremos 80% para treinamento e 20% para teste. Essa divisão quando estamos com dados de forma temporal, ocorre considerando os primeiros $x\%$ dos dados para treinamentos e em seguida os $100\% - x\%$ para teste, em forma cronológica.

Com a parte dos dados de treinamento, ensinamos o computador através de um modelo matemático a como fazer a previsão da variável resposta y, que é a variável que desejamos prever. Quando treinamos o modelo, passamos tanto a variável resposta y quanto as variáveis explicativas, denotadas matricialmente por X, isto com a finalidade do computador aprender como são as possíveis respostas ou previsões por meio de variáveis explicativas. Após o computador aprender a fazer as previsões do problema em questão, apresentamos as variáveis explicativas X separadas para testes, que são os $100\% - x\%$ dos dados para o computador prever a variável resposta y ou rótulos dos novos dados, usando o modelo matemático utilizado no treinamento, sem passar os rótulos ao modelo.

Por fim aplicamos medidas de avaliação para verificarmos se o modelo realizou as previsões para os dados de teste de forma correta e também com o intuito de sabermos qual a taxa de acerto do computador com o modelo utilizado.

5.3.1 Validação Cruzada

Quando aplicamos algum modelo de *Machine Learning* dependemos dos dados utilizados e dos parâmetros do modelo para obter uma boa previsão. Para isso precisamos testar um conjunto grande de valores para os hiperparâmetros de forma a escolher o que resulta em um melhor resultado no conjunto de treinamento. Esta seleção dos melhores hiperparâmetros é realizada utilizando a técnica conhecida por Validação Cruzada ou Cross-Validation (CV).

A Validação Cruzada (CV) é uma técnica que leva em consideração a divisão dos dados em k subconjuntos, de forma aleatória, e cada um dos modelos é treinado usando uma combinação distinta destes subconjuntos. Em cada uma das k divisões, é separada uma parte para teste, com essa parte avaliamos a qualidade da k-ésima divisão utilizando uma medida de avaliação [26], conforme vemos na Figura 5.2.

Para cada uma das k divisões, as n combinações dos parâmetros são realizadas. Por exemplo, se considerarmos um modelo com 3 hiperparâmetros e em cada um testarmos 4 possibilidades de valores utilizando Validação Cruzada com $k = 5$, teremos 320 modelos avaliados, ou seja

$$\underbrace{4}_{\text{opções } 1^{\text{o}} \text{ hiperparâmetro}} \times \underbrace{4}_{\text{opções } 2^{\text{o}} \text{ hiperparâmetro}} \times \underbrace{4}_{\text{opções } 3^{\text{o}} \text{ hiperparâmetro}} \times \underbrace{5}_{\text{quantidade de divisões CV}} .$$

No final, escolhemos a combinação de hiperparâmetros que resultar em melhor resultado em relação à medida de avaliação escolhida.

Após a escolha dos melhores hiperparâmetros com a técnica CV aplicamos no modelo esses hiperparâmetros com os dados de treinamento. E posteriormente avaliamos a

predição de novos dados com os dados de teste, que são os 20% dos dados que separamos antes de aplicar os métodos de Machine Learning.

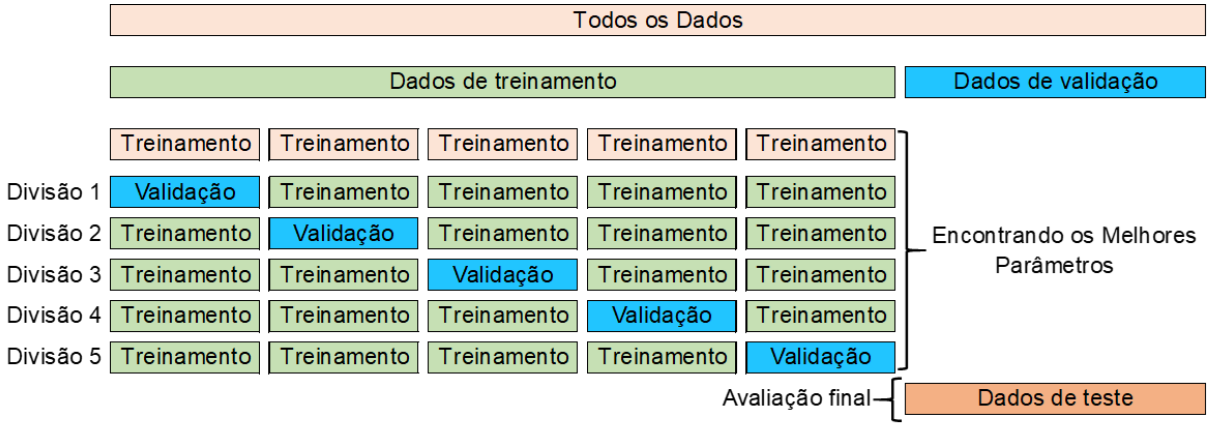

Figura 5.2: Validação Cruzada K-fold.
Fonte: Adaptado de [63].

Na realização de uma análise usando dados de forma temporal não podemos usar a divisão conforme mostrada anteriormente. Neste caso, a divisão ocorre de forma que os índices dos dados de teste sejam maiores que os dados de treinamento, conforme Figura 5.3. Ou seja, na primeira divisão treinamos usando o primeiro segmento para treinar e o segundo para fazer a validação com o conjunto de hiperparâmetros, na segunda divisão usamos os dois primeiros segmentos para treinamento e o terceiro para teste e assim sucessivamente até acabar os segmentos de dados [40]. Esta forma será a aplicada durante o trabalho, pois os dados financeiros têm uma sequência temporal nas datas.

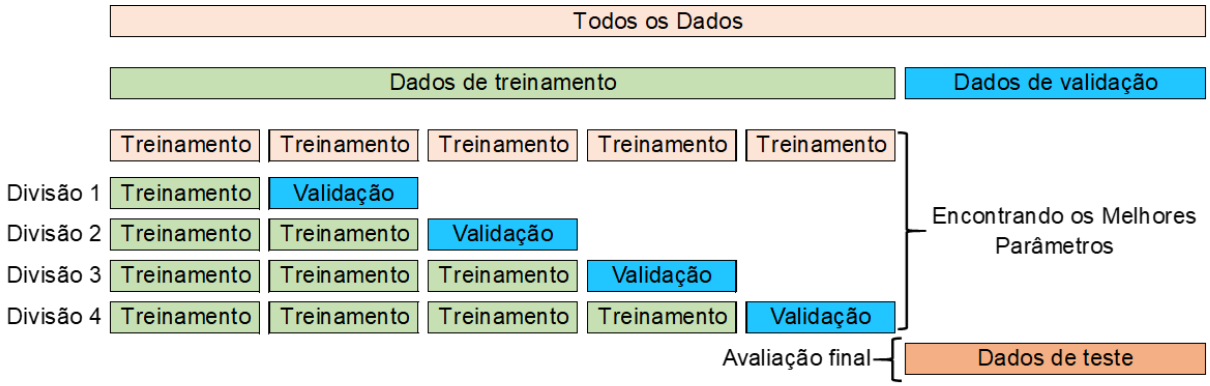

Figura 5.3: Validação Cruzada usando divisão temporal.
Fonte: Adaptado de [68].

Nas próximas seções, apresentaremos os modelos de ML aplicados ao problema de previsão da tendência dos índices IBOV e IFIX. Estes métodos são utilizados para ensinar ao computador como fazer a previsão dos rótulos no caso de problemas de classificação e valores reais quando estamos em um problema de regressão.

5.3.2 Regressão Logística

A Regressão Logística (RL) é um modelo empregado em problemas onde a ocorrência do evento acontece de forma binária, por exemplo, as classes da variável resposta são sim ou não, 0 ou 1, defeito ou não defeito, sucesso ou fracasso, etc. Logo, a RL é modelada com a distribuição Bernoulli.

O modelo de RL relaciona a probabilidade de um evento ocorrer com as variáveis explicativas (X), também denominadas por preditoras, $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$, $k \in \{1, \ldots, n\}$, $i \in \{1, \ldots, m\}$, em que i é o número de observações e k o número de variáveis [15], por meio da equação do modelo

$$transformação(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}. \tag{5.1}$$

Quando treinamos o modelo de RL estamos ensinando à máquina quais os valores de β que minimizam os erros associados às predições.

A partir da equação (5.1) do modelo, escolhemos uma transformação de maneira adequada de forma que a definição de função seja respeitada, e que exista uma lógica prática. Adotaremos a transformação logit que é comumente utilizada para p_i [15]. Doravante, a equação (5.1) do modelo fica reescrita da seguinte maneira

$$\begin{aligned}
logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \\
\implies \frac{p_i}{1-p_i} &= e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}} \\
\implies p_i(1 + e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}}) &= e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}} \\
\implies p_i &= \frac{e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}}}{1 + e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}}} \\
\implies p_i &= \frac{1}{1 + e^{-(\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i})}} \\
\implies p_i &= \frac{1}{1 + e^{-(X^T\beta)}},
\end{aligned} \tag{5.2}$$

onde $X^T = (1, x_{1,i}, x_{2,i}, \ldots, x_{k,i})$ e $\beta = (\beta_0, \ldots, \beta_k)$.

Após a aplicação da transformação, obtemos a equação (5.2) do modelo Regressão Logística, a qual tem a forma da função logística como pode ser vista na Figura 5.4.

Em certos problemas que estamos resolvendo, pode ocorrer sobreajuste (*overfitting*) ou subajuste (*underfitting*). Nos casos em que existe a presença de uma destas adversidades no ajuste dos dados usando RL aplicamos um termo adicional de regularização na minimização da função custo [39]. Existem três tipos de regularização: Ridge, Lasso e Elastic Net.

No geral, podemos resolver o modelo de regressão logística como um problema de regressão linear da forma

$$\min_{\beta} \sum_{j}^{k} (y_j - p_j)^2, \tag{5.3}$$

em que y_j é o valor real.

Na regularização Lasso, acrescentamos ao problema 5.3 o termo da soma dos módulos dos valores de todos os coeficientes do modelo [39]

$$\min_{\beta} \sum_{j}^{k} (y_j - p_j)^2 + \lambda \sum_{l}^{k} ||\beta_l||_1,$$

sendo λ o parâmetro que regula a penalidade, isto é, quanto maior mais penalizamos.

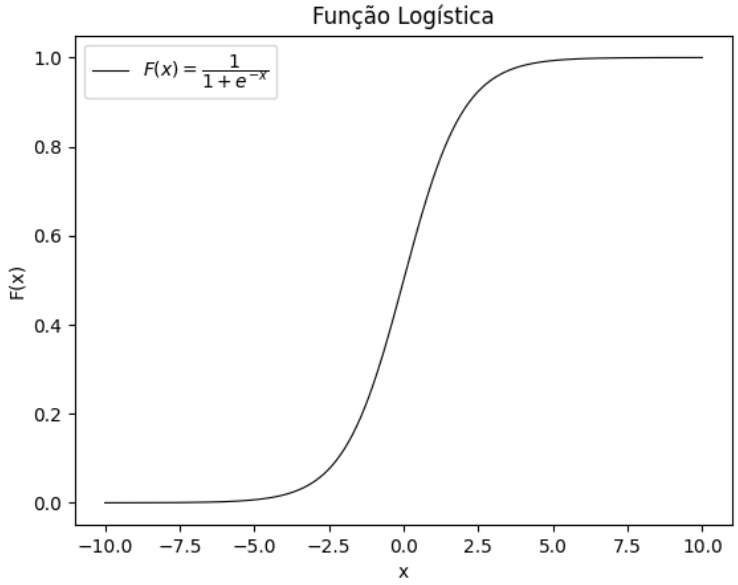

Figura 5.4: Função Logística.
Fonte: O Autor (2021).

Na regularização Ridge o termo adicional de penalidade acrescentado no problema 5.3 é soma dos coeficientes ao quadrado

$$\min_{\beta} \sum_{j}^{k} (y_j - p_j)^2 + \lambda \sum_{l}^{k} \beta_l^2.$$

Por último, na regularização Elastic Net combinamos os dois tipos de regularização anteriores, obtendo

$$\min_{\beta} \sum_{j}^{k} (y_j - p_j)^2 + \lambda_1 \sum_{l}^{k} ||\beta_l||_1 + \lambda_2 \sum_{l}^{k} \beta_l^2,$$

sendo que λ_1 e λ_2 são parâmetros que regulam a penalidade e podem ou não ter valores iguais.

5.3.3 Floresta Aleatória

No modelo de Floresta Aleatória, do inglês *Random Forest* (RF), no lugar de gerar uma única árvore através do modelo Árvore de Decisão criamos múltiplas árvores [1]. Para

classificar o novo conjunto de variáveis explicativas X numa classe da variável resposta (y), a floresta escolhe a classe com maior número de votos na etapa de votação conforme o fluxograma da Figura 5.5.

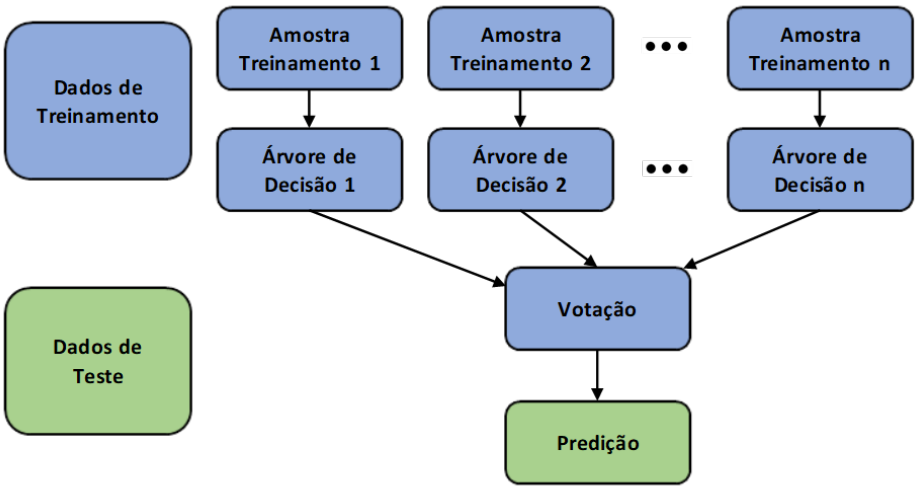

Figura 5.5: Fluxograma das etapas da aplicação do modelo Floresta Aleatória.
Fonte: Adaptado de [77].

Como usamos uma Árvore de Decisão para a criação das árvores no modelo RF, iremos explicar antes como funciona o modelo de Árvore de Decisão. A Árvore de Decisão é um modelo de aprendizagem supervisionada, em que dividimos os dados em dois ou mais grupos homogêneos com base nas variáveis explicativas [59]. Este modelo aceita tanto dados categóricos como numéricos, por isso não precisa de técnicas de representação para variáveis categóricas.

Em uma árvore temos o nó onde cada árvore é criada, contendo o nó raiz que separa no melhor grupo de dados, o nó de decisão usado para tomada de decisão para criar as novas árvores e o nó terminal onde termina a árvore e não é criada mais nenhuma outra [59], conforme mostrado na Figura 5.6.

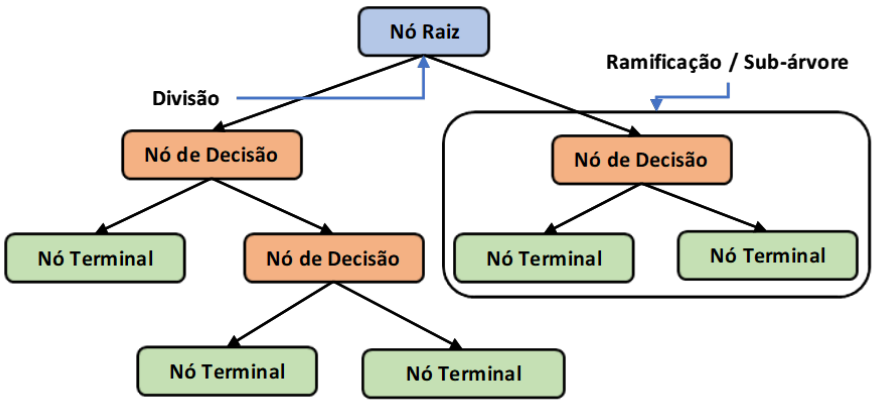

Figura 5.6: Exemplo de uma Árvore de Decisão.
Fonte: Adaptado de [1]

O objetivo da Árvore de Decisão é efetuar a separação dos dados da melhor forma possível, considerando o subconjunto com o valor superior da medida de impureza [59]. Para calcular o valor da medida de impureza usamos o índice de Gini, Entropia, entre outros. Segundo Neelam Tyagi [78] a entropia e o índice de Gini são calculados usando as equações a seguir, 5.4 e 5.5, respectivamente.

$$Entropia = -\sum_{i=1}^{k} p_i log_2(p_i), \tag{5.4}$$

sendo p_i a probabilidade de um evento i, $1 \leq i \leq k$, ocorrer.

$$\acute{I}ndice\ de\ Gini = 1 - \sum_{i=1}^{k} (p_i)^2, \tag{5.5}$$

em que p_i a probabilidade de um elemento ser classificado em uma classe distinta.

De acordo com [1], e conforme o fluxograma da Figura 5.5, as etapas para a geração de novas árvores em um modelo RF são as seguintes

1. Seja N o número de observações da base de treinamento. Desta base são escolhidas n amostras aleatoriamente com reposição. Estas amostras serão usadas para ajustar cada uma das árvores,

2. Das M variáveis, escolhemos um número m com $m \leq M$ de maneira que cada nó teste m variáveis aleatoriamente. A melhor separação nas m é utilizada para separar o nó em duas partes. Devemos manter m constante enquanto criamos mais árvores na floresta,

3. Cada uma das árvores é cultivada no maior tamanho permitido, sem ocorrer podas. A predição da classe de novos atributos é feita juntando as previsões das árvores usando o voto majoritário.

5.3.4 *XGBoost*

O modelo *Extreme Gradient Boost* também conhecido por *XGBoost*, foi elaborado por Chen e Guestrin em 2016 [12]. Segundo Morder [41], o *XGBoost* é uma evolução do modelo de Árvore de Decisão, passando por outros métodos conforme a sequência da lista a seguir.

1. **Árvore de Decisão**: representação gráfica de decisões fundamentadas em certas escolhas,

2. **Bagging**: é um *ensemble* com combinação de meta-algoritmo com múltiplas árvores de decisão com o processo de escolha pela votação majoritária,

3. **Floresta Aleatória**: algoritmo fundamentado em Bagging no qual um subconjunto das variáveis explicativas são selecionadas de maneira aleatória para construir cada árvore,

4. **Boosting**: modelos produzidos de forma sequencial para minimizar os erros dos modelos anteriores,

5. **Gradient Boosting**: emprega o Método do Gradiente para minimizar os erros dos modelos sequenciais,

6. ***XGBoost***: é o modelo do Gradient Boosting otimizado através de processamento paralelo que poda as árvores, trata os valores em falta e aplica regularização para evitar overfitting.

As melhorias no modelo *XGBoost* em relação ao do Gradient Boosting em [41] e [12] são as seguintes

- uso de regularização nos modelos mais complexos, adicionando um termo a mais na minimização da função objetivo como descrito na Seção 5.3.2,
- no XGBoost se considera dados esparsos nos dados de treinamento, para realizar o aprendizado do melhor valor faltante, consequentemente lidando mais eficientemente com padrões de dispersão,
- o algoritmo utiliza na implementação o "Weighted Quantile Sketch"[12], ou algoritmo do esboço de quantil ponderado, para encontrar as divisões ideais entre os conjuntos de dados ponderados,
- por último, no algoritmo do modelo *XGBoost* já é apresentada uma Validação Cruzada em cada iteração.

Uma das características principais desse modelo é o treinamento muito mais rápido quando comparado com os modelos predecessores.

5.3.5 Máquina de Vetores Suporte

A Máquina de Vetores Suporte, ou Support Vector Machine (SVM), representa uma técnica supervisionada de aprendizagem de máquinas, da mesma forma que outros modelos apresentados neste trabalho. O modelo SVM funciona muito bem para problemas de classificação; já sua versão para regressão tem, em geral, um desempenho inferior comparado a outros métodos.

No caso linear, o objetivo do SVM é obter uma curva que divida o espaço em dois hiperplanos que separe os dados da melhor forma possível [31]. No problema de classificação binária, considerando as classes dadas pelos rótulos $y_i = -1$ e $y_i = 1$, $i \in (1, \ldots, n)$, um hiperplano é definido pela equação (5.6) [57]

$$w^T x + b = 0, \tag{5.6}$$

em que w é um vetor de pesos e $b \in \mathbb{R}$ é o intercepto, ao passo que para que cada ponto estar do lado correto em relação à curva, é necessário que cada ponto satisfaça

$$\begin{cases} w^T x + b < -1, \text{ para } y_i = -1 \\ w^T x + b \geq 1, \text{ para } y_i = +1. \end{cases}$$

Consequentemente, encontramos a margem otimizada que é o intervalo que compreende do(s) vetor(es) suporte de uma classe ao(s) da outra classe, para dado peso w e intercepto b. Os vetores de suporte considerados são os pontos mais próximos de cada

uma das classes ao hiperplano. Neste caso estamos com SVM com margens rígidas. Os elementos do modelo SVM podem ser vistos na Figura 5.7.

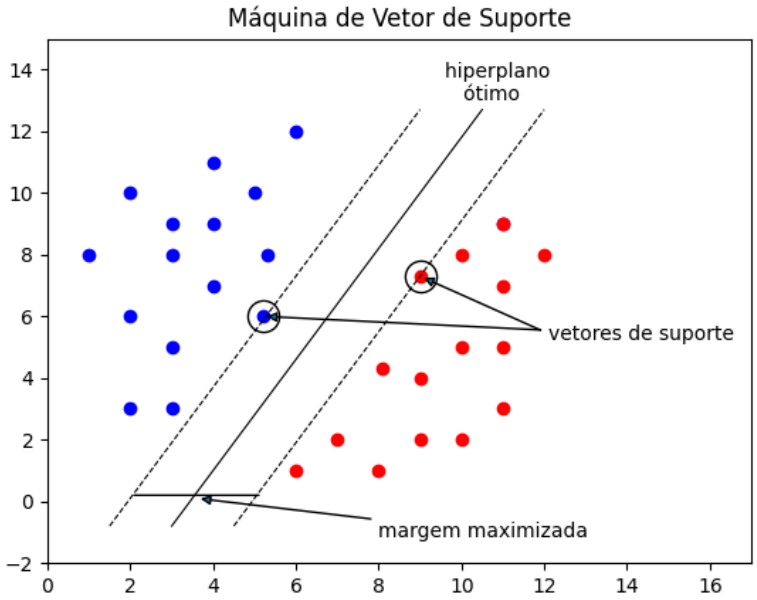

Figura 5.7: Exemplo do modelo SVM linearmente separável.
Fonte: O Autor (2021).

Como desejamos obter a maximização da margem, respeitando a equação (5.6), obtemos isso através do problema de minimização (5.7) [5].

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s. a.} \quad & y_i(w^T x_i + b) \geq 1, \text{para todo } i = 1, 2, \ldots, m. \end{aligned} \tag{5.7}$$

Em alguns casos, não conseguimos separar linearmente os dados através de um hiperplano como realizado até o momento. Este caso é denominado por SVM não linear ou SVM com dados linearmente não separáveis, como o exemplo da Figura 5.8.

Conforme Meloni [38], para conseguirmos mapear dados linearmente não separáveis devemos relaxar as restrições do problema analisado, que considerava critérios rígidos. Ou seja, suavizamos a equação (5.3.5) aplicando uma variável de folga $\xi_i \geq 0$ para todo i, $i = 1, \ldots, n$, obtendo assim as condições

$$\begin{cases} w^T x + b < -1 + \xi_i, \text{ para } y_i = -1 \\ w^T x + b \geq 1 - \xi_i, \text{ para } y_i = +1. \end{cases} \tag{5.8}$$

Doravante, a otimização apresentada na equação (5.7) aplicando a suavização resultará na otimização dada na equação (5.9) [5].

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{M} \xi_i \\ \text{s. a.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, m \\ & \xi_i \geq 0 \ \forall i, \end{aligned} \tag{5.9}$$

sendo C o parâmetro de penalização de margem.

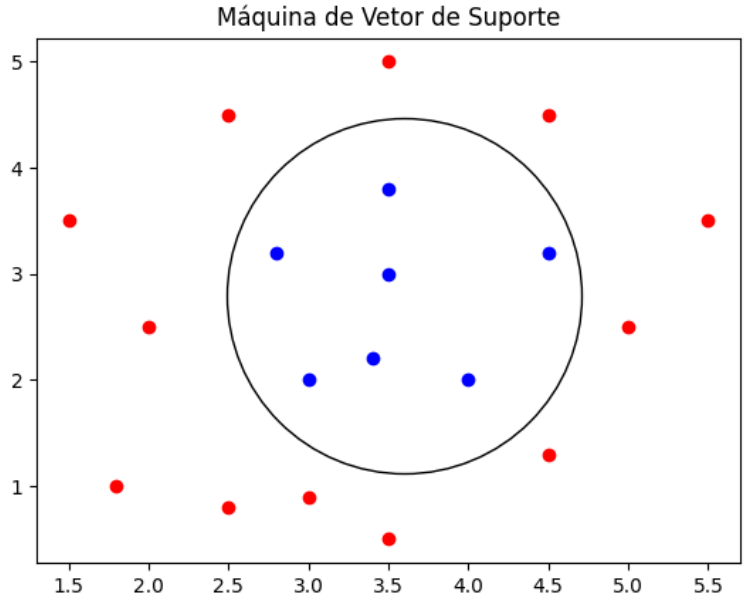

Figura 5.8: Exemplo do modelo SVM linearmente não separável.
Fonte: O Autor (2021).

O uso do conceito de núcleo, ou Kernel, parte do princípio que o espaço original seja mapeável em outro espaço de dimensão mais alta na motivação de que seja mais fácil separar os dados nesta dimensão, sem necessidade do emprego de muito recurso computacional [44], em conformidade com a Figura 5.9.

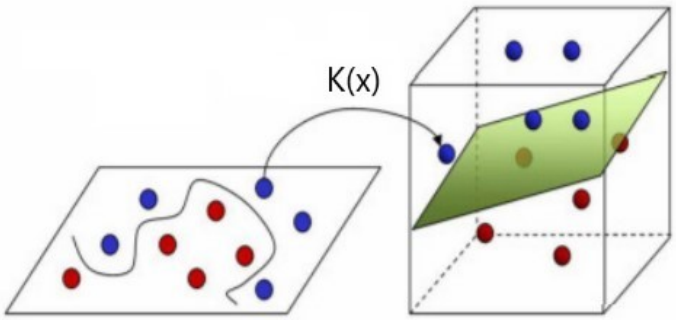

Figura 5.9: Conceito de Kernel com separação dos dados em dimensão superior.
Fonte: [49].

Tabela 5.1: Os núcleos mais comuns aplicados no modelo SVM.

Nome	**Kernel**
Linear	$K(x) = \langle x, x' \rangle$
Polinomial	$K(x) = (\gamma \langle x, x' \rangle + r)^d$
RBF	$K(x) = exp(-\gamma \|x - x'\|)^2$

Continua na próxima página

Nome	Kernel
Sigmoide	$K(x) = tanh\left(\gamma \left\langle x, x' \right\rangle + r\right)$

Fonte: Baseado em [62].

Os núcleos ou kernel's mais comuns, de acordo com [62], são os apresentados na Tabela 5.1. Os parâmetros presentes nos núcleos são: d o grau do polinômio, γ (positivo) é raio de influência de um elemento ou linha de dado no modelo como vetores de suporte, sendo valores grandes considerados longe e valores pequenos perto e r uma variável de controle.

5.3.6 Redes Neurais

O modelo de Rede Neural Artificial (RNA), ou Rede Neural (RN), foi proposto por um psiquiatra em parceria com um matemático em 1943 [35]. O modelo apresentado possui o objetivo de imitar o funcionamento das redes neurais naturais do sistema nervoso central para reconhecer padrões. Já em 1958 Rosenblatt [56], apresentou o primeiro modelo implementado de RNA designado de Rede Neural perceptron. Os modelos de RN têm sido utilizados em maior quantidade e escala apenas há algumas décadas, devido à melhora do processamento de computadores.

O neurônio artificial, designado por perceptron, tem referência ao neurônio natural representado na Figura 5.10. O perceptron calcula a saída do modelo, quando são dadas n entradas, as quais contém os dados das variáveis explicativas (x_i, $i \in \{1, \ldots, m\}$) e a variável resposta y [29], isto é, em termo computacional dado n *inputs* obtemos um *output*.

A Rede Neural Perceptron Multi-Camadas (RN MLP) é uma RN composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída tendo seus respectivos pesos $w_i, i = \{1, \ldots n\}$ [71], conforme Figura 5.11. A resposta de cada camada é passada para a camada subsequente até chegar à saída (*output*), como retratado na Figura 5.11.

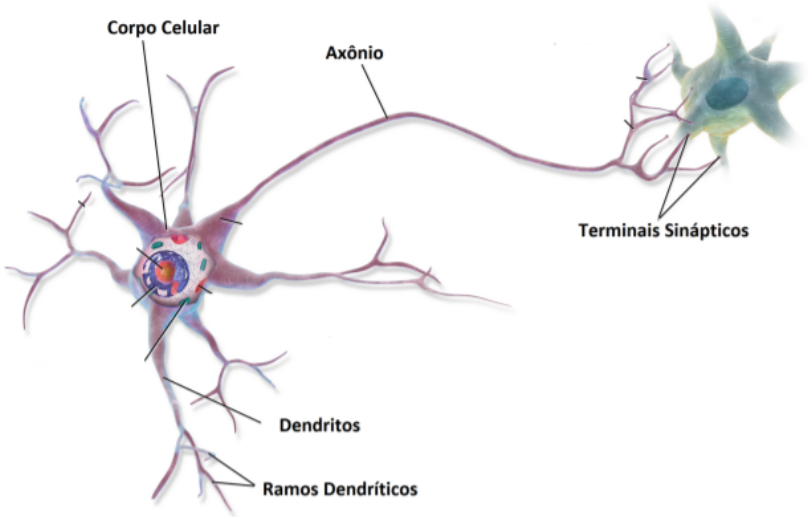

Figura 5.10: Representação das partes que compõem um neurônio do cérebro.
Fonte: Adaptado de [26].

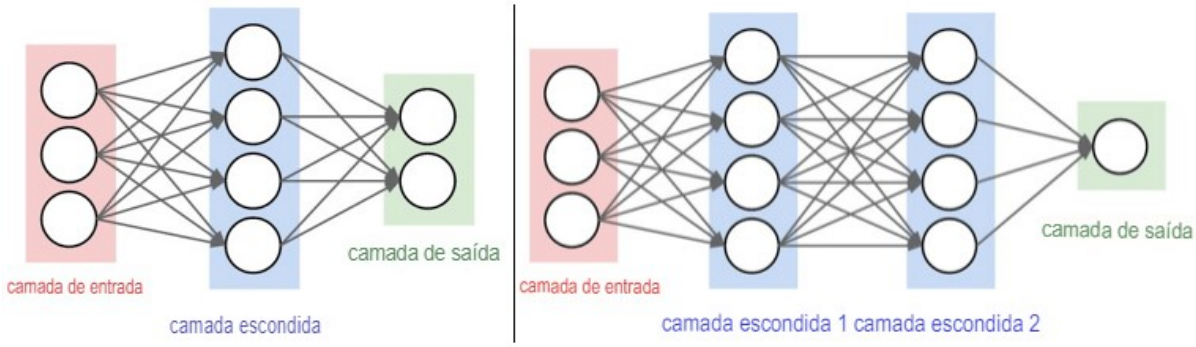

Figura 5.11: Representação de uma Rede Neural Artificial.
Fonte: Adaptado de [37].

Inicialmente, a sugestão de Rosenblatt foi utilizarmos o modelo com a saída do modelo binária, onde consideramos um limite dado pelo usuário que depende do problema analisado, como mostrado na equação (5.10) [29]. Desta forma, a tomada de decisão que tomamos é baseada na importância de cada uma das variáveis da base de dados utilizadas na entrada, ou seja,

$$output = \begin{cases} 0, z \leq limite \\ 1, z > limite \end{cases}, \tag{5.10}$$

em que, $z = \sum_{i=1}^{n} w_i x_i$, $i = \{1, \ldots, n\}$, sendo x_i a variável de entrada e w_i o peso conferido à variável.

Atualmente, a utilização do modelo RN pelo computador ocorre definindo vetorialmente as variáveis [29], em consonância com a expressão

$$z = \sigma(wx + b),$$

sendo σ uma função de ativação qualquer.

A função de ativação determina se um neurônio é ativado ou não utilizando o esquema da equação (5.10), sendo que caso a soma seja maior que um valor então o neurônio é ativado, caso contrário é desativado [71]. Esta função pode ser linear ou não linear. Na Tabela 5.2 apresentamos as principais funções de ativação. Contudo, podemos criar nossa própria função de ativação, que podem ou não trazer resultados relevantes. Os gráficos destas funções de ativação podem ser vistos na Figura 5.12.

Tabela 5.2: Algumas das função de ativação que podem ser utilizadas em Redes Neurais.

Nome	**Função de Ativação**
Linear	$\sigma(z) = z$
ReLU	$\sigma(z) = \max(0, z)$
Sigmoide	$\sigma(z) = \frac{1}{1+e^{-z}}$
TanH	$\sigma(z) = \frac{2}{1+e^{-2z}} - 1$

Continua na próxima página

Nome	Função de Ativação
Softmax	$\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}}, j = \{1, \ldots, k\}$
Swish	$\sigma(z) = \frac{z}{1 + e^{-z}}$

Fonte: Baseado em [26] e [34].

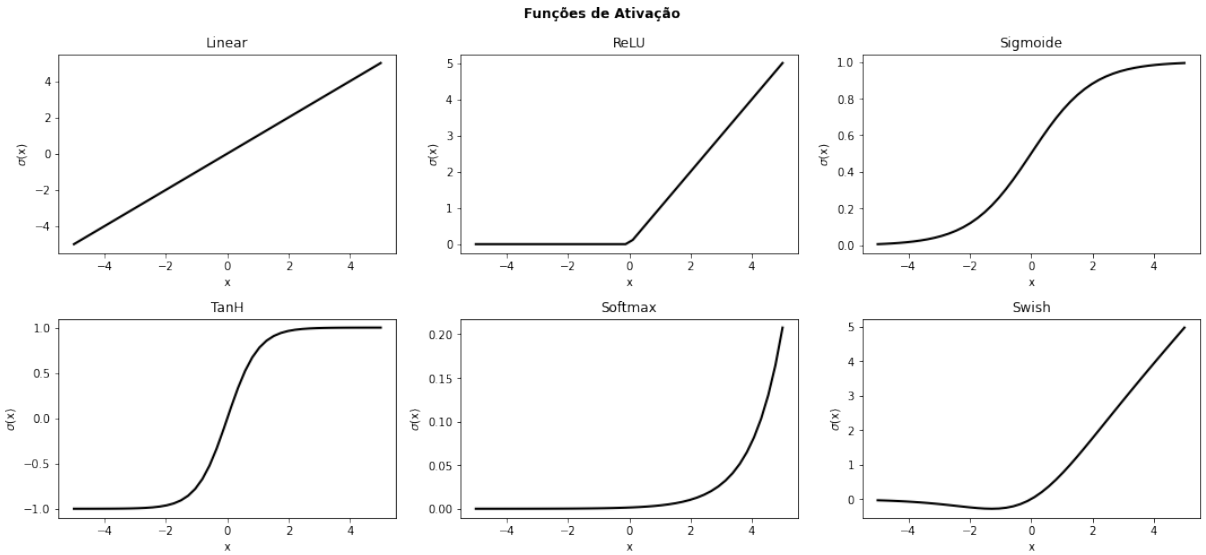

Figura 5.12: Funções de ativação usadas no modelo RN.
Fonte: O Autor (2021).

5.3.7 Medidas de Avaliação dos Modelos de ML

Para avaliação dos modelos de *Machine Learning* usamos algumas medidas obtidas através da matriz de confusão. Considere C a matriz de confusão, em que cada $C_{i,j}$ é o número de observações conhecidas ser do grupo i e predita pelo grupo j [66]. Em outros termos, por meio da matriz de confusão dispomos da combinação entre os valores reais e as previsões obtidas na aplicação do modelo treinado pelo computador nos dados de teste, conforme Figura 5.13. Essa avaliação serve para examinarmos o quanto o modelo teve ou não de assertividade nas previsões.

		Valor Predito	
		Não (0)	**Sim (1)**
Valor Real	**Não (0)**	Verdadeiros Negativos (VN)	Falsos Positivos (FP)
	Sim (1)	Falsos Negativos (FN)	Verdadeiros Positivos (VP)

Figura 5.13: Matriz de confusão do modelo com duas classes.
Fonte: Adaptado de [42].

Da matriz de confusão podemos obter os seguintes valores:

- **Verdadeiros Negativos (VN)**: número de observações da classe negativa que foram preditas como sendo da classe negativa,

- **Verdadeiros Positivos (VP)**: número de observações da classe positiva que foram preditas como sendo da classe positiva,
- **Falsos Positivos (FP)**: números de observações preditas como positivas, mas que são da classe negativa,
- **Falsos Negativos (FN)**: números de observações preditas como negativas, mas que são da classe positiva.

As medidas de avaliação obtidas a partir da matriz de confusão são as seguintes [42]:

- **Precisão**: porcentagem de valores da classe positiva predita corretamente

$$PREC = \frac{VP}{VP + FP},$$

- **Recall**: porcentagem de valores da classe positiva predito corretamente dentre o total de valores da classe sim que deveriam ser classificadas corretamente

$$RC = \frac{VP}{VP + FN},$$

- **F1-Score**: é a média harmônica entre a precisão e o recall

$$f1_{score} = \frac{2}{\frac{1}{PREC} + \frac{1}{RC}} = \frac{2 \cdot PREC \cdot RC}{PREC + RC} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN},$$

- **Acurácia**: porcentagem de valores preditos corretamente tanto da classe negativa quanto da classe positiva

$$ACC = \frac{VN + VP}{VN + VP + FN + FP},$$

- **Acurácia Balanceada**: é a média do recall obtida em cada classe

$$ACC_{bal} = \frac{\frac{VN}{VN+FP} + \frac{VP}{VP+FN}}{n},$$

em que n é o número de classes.

As medidas de avaliação Acurácia Balanceada, Recall e F1-Score são bastante empregadas em problemas em que as classes são desbalanceadas.

5.3.8 Análise de Variância (ANOVA)

Com o intuito de verificarmos quais as variáveis explicativas mais significativas entre as disponíveis, fazemos o uso da ANOVA. Aplicando a ANOVA podemos averiguar se as variáveis são dependentes ou não uma em relação às demais [61]. As tabelas das variáveis mais significativas podem ser vistas no Anexo C.

ANOVA (*Analysis Of Variance*) é utilizada para verificarmos se existe diferença significativa entre as médias dos dados de dois ou mais grupos [73]. Para a ANOVA as hipóteses utilizadas são

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \\ H_1 : \mu_i \neq \mu_j, \end{cases}$$

em que $i \neq j$ e k é o número de amostras.

Na ANOVA consideramos duas fontes de variação, uma entre grupos e outra dentro dos grupos, conforme a Tabela 5.3 [73]. A soma de quadrados totais é dada por

$$SQ_{total} = SQ_{entre} + SQ_{dentro}$$
$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2,$$

onde

- y_{ij} é a observação da repetição j dentro da amostra i,
- $\bar{y}_{i.}$ são as médias das amostras y_{ij},
- $\bar{y}_{..}$ média geral de y_{ij}.

Tabela 5.3: Tabela da análise de variância

Fonte de Variação	**Graus de Liberdade** (gl)	**Soma de Quadrados**	**Quadrado Médio**	**F calculado** (F_{calc})
Entre	$k-1$	SQ_{entre}	$QM_{entre} = SQ_{entre}/gl$	$\frac{QM_{entre}}{QM_{dentro}}$
Dentro	$n-k$	SQ_{dentro}	$QM_{dentro} = SQ_{dentro}/gl$	
Total	$n-1$	SQ_{total}		

Fonte: Adaptado de [73].

Com a tabela da ANOVA, a tomada de decisão que teremos é que se $F_{k-1,n-k} > F_{calc}$ aceitamos a hipótese nula H_0.

Para utilizarmos a ANOVA para seleção de variáveis seguimos os seguintes passos na linguagem de programação Python, biblioteca scikit-learn:

1. Utilizamos a função 'f_classif' [65] para calcular o valor F_{calc}, de cada uma das variáveis explicativas, através da ANOVA. A ANOVA é realizada entre cada uma das variáveis explicativas e a variável resposta.

2. Selecionamos o percentil de x% das variáveis mais significativas dentre todas as variáveis que utilizamos em cada uma das análises de tendência, usando a função 'SelectPercentile' [64], também da biblioteca scikit-learn. Essa função considera os valores de F_{calc} obtidos através da ANOVA no passo 1.

Assim, com o valor F (*F Score*) obtido a partir da ANOVA, verificamos quais as variáveis mais significativas, considerando que quanto maior o valor F mais indícios temos que as variáveis diferem entre si, o inverso vale para p-valor [61].

5.4 Teste de Hipóteses

Nas análises comparativas entre os índices é usada como metodologia a aplicação de testes estatísticos (paramétricos ou não paramétricos) para verificarmos a igualdade de média para dados emparelhados e igualdade de variância (homocedasticidade). Antes de usar os métodos, verificamos se o ideal em cada caso é usar método paramétrico ou não paramétrico, para isso utilizamos o teste de normalidade Kolmogorov-Smirnov (KS) [70]. Se os dados mostrarem ser normalmente distribuídos, segundo o teste KS, utilizamos testes paramétricos, caso contrário os testes não paramétricos.

5.4.1 Teste de Normalidade Kolmogorov-Smirnov

Na maior parte dos métodos estatísticos existe a hipótese de que os dados são gerados com uma distribuição de probabilidade específica. No caso da distribuição de probabilidade normal, o teste de Kolmogorov-Smirnov (KS) é utilizado para verificar se os dados são normalmente distribuídos, $X_i \sim N(\mu, \sigma^2)$, $i \in \{1, \ldots, n\}$. Contudo o teste de hipótese KS, pode ser utilizado para verificar qualquer distribuição de probabilidade.

O teste KS corresponde a comparação entre a distribuição de frequência acumulada observada, obtida a partir dos dados, e a distribuição de frequência acumulada decorrente da distribuição teórica [70]. Considere $S(x_i)$ a distribuição de frequências relativas acumuladas da distribuição normal e $S_n(x)$ a distribuição de frequências relativas acumuladas observadas da variável aleatória com n observações. Se x_i é um valor possível qualquer, então $S_n(x_i) = \dfrac{S_i}{n}$, em que S_i é a proporção esperada de observações menores ou iguais a x_i. Neste teste avaliamos as hipóteses

$$\begin{cases} H_0 : \text{Os dados seguem uma distribuição normal, isto é, } S_n(x) = S(x)\ \forall x \\ H_1 : \text{Os dados não seguem uma distribuição normal, isto é, } S_n(x) \neq S(x)\ \forall x \end{cases},$$

considerando como um teste bilateral.

Durante o teste KS o valor calculado considerado, que comparamos com o valor crítico, é dado pelo máximo dos valores absolutos das diferenças $S(x_i) - S_n(x_i)$, denominado por desvio máximo, ou seja,

$$D_{calc} = max|S(x_i) - S_n(x_i)|.$$

Sob H_0 a estatística do teste segue distribuição de Kolmogorov-Smirnov com nível de significância α e os passos para aplicarmos o teste de normalidade KS são os seguintes:

1. Calcular $S(x_i)$, encontrada na tabela da distribuição normal usando a transformação $Z_i = \dfrac{x_i - \bar{x}}{s}$, sendo $\bar{x}$ a média dos dados e s a variância dos dados,
2. Calcular $S_n(x_i)$,
3. Computar o valor calculado D_{calc},
4. Procurar o valor crítico D_{crit} na tabela da distribuição KS usando o tamanho da amostra n e o nível de significância α, para o caso de $n \leq 40$. Caso $n > 40$, utilizamos o seguinte valor crítico $D_{crit} = \dfrac{1.36}{\sqrt{n}}$,

5. Comparar o valor calculado D_{calc} e o valor crítico D_{crit} para saber se aceitamos ou rejeitamos a hipótese nula H_0. Aceitamos H_0 quando $D_{calc} < D_{crit}$.

5.4.2 Teste de Igualdade de Média Pareado (Teste T)

O teste de igualdade de média é usado para verificarmos se a média das duas amostras é significantemente igual. Para a utilização de testes de igualde de média devemos considerar, amostras de populações independentes. À vista disto, usamos o teste t para dados pareados. Os dados são pareados quando os elementos de uma amostra ou grupo têm uma correlação entre si e cada amostra tem um par com a outra amostra. Em nosso estudo, cada amostra de um índice tem correlação entre si e possui um par com outro índice tendo à data em comum. Assim, podemos usar esse teste de igualdade de média pareado.

Sejam $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ duas amostras dependentes. Como as observações são pareadas, temos os pares de amostras (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) e consequentemente, definimos $D_i = X_i - Y_i, i = 1, 2, \ldots, n$, a diferença entre os elementos de cada par. Além disso, consideramos que $D_i \sim N(\mu_D, \sigma_D^2)$ [48], isto é, as diferenças D_i são normalmente distribuídas com média μ_D e variância σ_D^2. Na realização deste teste consideramos uma das hipóteses:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}, \ ou \ \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D > 0 \end{cases}, \ ou \ \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D < 0 \end{cases},$$

as quais são denominadas por bilateral, unicaudal à direita e unicaudal à esquerda, respectivamente.

O teste será realizado utilizando o valor calculado de t seguinte

$$t_{calc} = \frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}},$$

em que a μ_D é estimado pela média das diferenças das amostras e a variância da amostra é calcula por $s_D^2 = \dfrac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}$. A hipótese H_0 segue distribuição t de *Student* com $n-1$ graus de liberdade. Para a utilização do teste de igualdade de média para dados pareados, devemos seguir os seguintes passos:

1. Escolher entre uma das hipóteses: bilateral, unilateral à direita e unilateral à esquerda,

2. Calcular o valor de t_{calc} sob a hipótese nula,

3. Procurar o valor crítico de t (t_{crit}) na tabela da distribuição t de *Student* com $n-1$ graus de liberdade e nível de significância α, conforme a hipótese considerada no passo 1:

   (i) Bilateral: $-t_{\alpha/2}$ e $t_{\alpha/2}$,

   (ii) Unilateral à direita: t_α,

   (iii) Unilateral à esquerda: $-t_\alpha$,

4. Comparar t_{calc} e t_{crit} para sabermos se aceitamos ou rejeitamos H_0. Para tal, aceitamos a hipótese H_0 quando:

   (i) Bilateral: $-t_{\alpha/2} < t_{calc} < t_{\alpha/2}$,

   (ii) Unilateral à direita: $t_{calc} < t_{\alpha}$,

   (iii) Unilateral à esquerda: $t_{calc} > -t_{\alpha}$.

5.4.3 Teste de Igualdade de Média Pareado (Teste Wilcoxon)

O teste Wilcoxon pareado é um teste alternativo ao teste t para quando os dados não possuem estatisticamente a distribuição normal, ou melhor dizendo, estamos no caso de um teste não paramétrico. Este teste é usado para fazermos a comparação da medida de posição de duas amostras, e sabermos se estas medidas de tendência central são estatisticamente iguais, no caso de dados pareados.

Sejam $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ duas amostras dependentes, de duas populações A_1 e A_2. Como as amostras das populações A_1 e A_2 são pareadas, temos os pares de amostras (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n). A partir disso, definimos $D_i = X_i - Y_i, i = 1, 2, \ldots, n$, a diferença entre os elementos de cada par [45]. Na aplicação do teste de Wilcoxon consideramos uma das hipóteses:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}, \ ou \ \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D > 0 \end{cases}, \ ou \ \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D < 0 \end{cases},$$

as quais são denominadas por bilateral, unicaudal à direita e unicaudal à esquerda, respectivamente.

Considere o posto das diferenças $|D_i|$, T_+ a soma dos postos com D_i positivos e T_- a soma dos postos em que $D_i \leq 0$, se haver postos iguais usamos a média entre os postos. Seja $T = min(T_-, T_+)$ [27]. Quando o tamanho da amostra for menor que 25, $n < 25$, consideramos as hipóteses acima. Porém, quando estamos com uma amostra grande, $n \geq 25$, T tem aproximadamente distribuição normal, e consideramos

$$\mu_T = \frac{n(n+1)}{4} \text{ e } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

A partir da média μ_T e desvio padrão σ_T, calculamos a estatística $T_{calc} = z = \dfrac{T - \mu_T}{\sigma_T}$ e comparamos com o valor da tabela da distribuição normal z (normal padrão).

Por fim, para saber se aceitamos ou rejeitamos a hipótese nula (H_0) comparamos T com T_{crit} ou, T com T_{calc} com z_{α} ou $z_{\alpha/2}$. Consequentemente, para $n > 25$ aceitamos a hipótese H_0 quando

(i) Bilateral: $-z_{\alpha/2} < T_{calc} = z < z_{\alpha/2}$,

(ii) Unilateral à direita: $T_{calc} = z < z_{\alpha}$,

(iii) Unilateral à esquerda: $-z_{\alpha} < T_{calc} = z$,

e para $n \leq 25$ quando $T > T_{crit}$, com T_{crit} dado na tabela do teste Wilcoxon.

5.4.4 Teste de Igualdade de Variância (Teste F)

O teste de igualdade de variância, também denominado por teste de homogeneiedade de variâncias, serve para verificarmos se duas amostras possuem variâncias significantemente iguais. Mas, especificamente o teste F é usado para fazermos a comparação da variância de duas populações normais independentes, σ_1^2 e σ_2^2 [47].

Seja $X_1, X_2, \ldots, X_{n_1}$ uma amostra com distribuição normal $N(\mu_1, \sigma_1^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ outra amostra com distribuição normal $N(\mu_2, \sigma_2^2)$, oriundas de populações diferentes. Na execução deste teste consideramos uma das hipóteses:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}, \quad ou \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{cases}, \quad ou \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases},$$

as quais são denominadas por bilateral, unicaudal à direita e unicaudal à esquerda, respectivamente.

A hipótese H_0 segue distribuição F com $n_1 - 1$ graus de liberdade para o numerador e $n_2 - 1$ graus de liberdade para o denominador, isto é, H_0 tem distribuição $F_{(n_1-1;n_2-1)}$. Os passos necessários para realização do teste F são os seguintes:

1. Escolher entre uma das hipóteses: bilateral, unilateral à direita e unilateral à esquerda,

2. Selecionar o nível de significância α,

3. Determinar a região crítica: precisamos procurar o valor crítico de F (F_{crit}), ou seja, procurar $F_{(n_1-1;n_2-1)}$ na tabela da distribuição F, conforme a hipótese considerada no passo 1:

   (i) Bilateral: $F_{\alpha/2}$ e $F_{1-\alpha/2}$,

   (ii) Unilateral à direita: $F_{1-\alpha}$,

   (iii) Unilateral à esquerda: F_{α},

4. Calcular o valor de F_{calc} sob a hipótese nula,

$$F_{calc} = \frac{S_1^2}{S_2^2},$$

5. Comparar F_{calc} e F_{crit} para sabermos se aceitamos ou rejeitamos H_0. Para tal, aceitamos a hipótese H_0 quando:

   (i) Bilateral: $F_{\alpha/2} < F_{calc} < F_{1-\alpha/2}$,

   (ii) Unilateral à direita: $F_{calc} < F_{1-\alpha}$,

   (iii) Unilateral à esquerda: $F_{calc} > F_{\alpha}$.

5.4.5 Teste de Igualdade de Variância (Levene)

O teste Levene é um teste não paramétrico para comparação de variâncias ou teste de homogeneidade para dados que não possuem o pressuposto de normalidade dos dados. Este teste foi elaborado por Levene em 1960 [33], e existem duas versões propostas por Brown e Forsythe em 1974, que são extensões que consideram no lugar da média da amostra na estatística do teste, a mediana em uma versão e a média aparada em outra versão [6]. Neste trabalho, consideramos a versão original. Para tal teste as hipóteses que analisamos são

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \\ H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para cada par } (i,j) \end{cases},$$

em que k é o número de amostras.

Também podemos notar que o teste de Levene pode ser considerado como uma variação da análise de variância (ANOVA) para k grupos, em que há uma transformação precedente nos dados [43]. Esta modificação anterior nos dados mede a distância absoluta entre cada um dos dados e a média do grupo.

Sejam $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ duas amostras que não possuem distribuição normal, isto é, $X \nsim N(\mu_1, \sigma_1^2)$ e $Y \nsim N\left(\mu_2, \sigma_2^2\right)$ [28]. A estatística do teste ou valor calculado é obtida pela expressão

$$L_{\text{calc}} = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i \left(\overline{Z}_{\text{i.}} - \overline{Z}_{\text{..}}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(Z_{\text{ij}} - \overline{Z}_{\text{i.}}\right)^2},$$

sendo

- N o total de amostras,
- N_i o tamanho da amostra i,
- $Z_{\text{ij}} = |Y_{\text{ij}} - \overline{Y}_{\text{i.}}|$, em que $\overline{Y}_{\text{i.}}$ pode ser a média da i-ésima amostra, a mediana da i-ésima amostra ou a média aparada de 10%,
- Y_{ij} cada replicação dentro da amostra i,
- $\overline{Z}_{\text{i.}}$ são as médias das amostras Z_{ij}, para $i \in \{1, \ldots, I\}$,
- $\overline{Z}_{\text{..}}$ é a média geral de Z_{ij}, para $i \in \{1, \ldots, I\}$ e $j \in \{1, \ldots, J\}$.

Em seguida, fazemos a comparação da estatística do teste (L_{calc}) com os valores críticos para concluirmos se rejeitamos ou aceitamos a hipótese nula (H_0). Desta forma, aceitamos a hipótese H_0 se $F_{(\alpha/2;n_1-1;n_2-1)} < L_{\text{calc}} < F_{(1-\alpha/2;n_1-1;n_2-1)}$.

Como no teste F, podemos encontrar o valor crítico de F (F_{crit}) na tabela da distribuição F. Também, conseguimos utilizar o teste de hipótese com as variações para os casos: unilateral à esquerda e unilateral à direita, além do caso bilateral considerado.

Capítulo 6

Análise de Tendência dos Índices Ibovespa e IFIX

Neste capítulo, realizamos uma análise baseada na dissertação de mestrado de Finkler [21], cujo propósito é fazer a predição da tendência de subida ou descida dos índices Ibovespa (IBOV) e Fundos de Investimentos Imobiliários (IFIX) utilizando dados mensais. Nesta análise o período de dados utilizado foi de Janeiro de 2001 a Maio de 2021 para o índice IBOV e de Janeiro de 2011 a Maio de 2021 para o índice IFIX. Exibiremos os principais resultados obtidos na análise.

6.1 Dados Utilizados

Seja t_i, $i = \{1, \dots N\}$, o valor do índice específico IBOV ou IFIX no i-ésimo mês e considere o horizonte $h \in \{1, 3, 6, 12\}$ para todo $i \in [4, N-h] \cap \mathbb{N}$. Pretendemos prever se o índice respectivo irá subir ou descer após h meses. Para isso, aplicamos os métodos de Aprendizagem de Máquinas apresentados na Seção 5.3. As variáveis explicativas X incluem dados do índice em questão, dados de mercado e variáveis obtidas através de *feature engineering*. Dentre estas variáveis, tanto o índice quanto cada uma das variáveis de mercado tem os valores [21]

$$t_i, t_{i-1}, t_{i-2}, t_{i-3},$$

em que desejamos prever o sinal do mês i, isto é, $y^{(i)} = sinal(t_{i+h} - t_i)$.

Os dados de mercado são os referentes à cotação do dólar, à cotação do ouro e ao Índice Nacional de Preços ao Consumidor Amplo (IPCA). Executamos duas análises: uma utilizando o índice IBOV (AIB) e outra aplicando o índice IFIX (AII). Os dados de mercado que consideramos na AIB são a cotação do dólar, a cotação do ouro e o índice IPCA. Já na análise AII os dados de mercado considerados foram os índices IBOV e IPCA.

Em relação à *feature engineering*, temos oito variáveis explicativas que são as seguintes

- **variável 1**: (mês 3 - mês 1) / mês 1;
- **variável 2**: coeficiente angular do índice IBOV para AIB e índice IFIX para AII;

- **variável 3**: coeficiente angular da cotação do dólar para AIB e do índice IBOV para AII;
- **variável 4**: coeficiente angular da cotação do ouro para AIB e do índice IPCA para AII;
- **variável 5**: coeficiente angular da cotação do IPCA para AIB, não utilizado na AII;
- **variável 6**: variável 2 / variável 3;
- **variável 7**: variável 2 / variável 4;
- **variável 8**: variável 2 / variável 5 para AIB, não utilizado na AII;

sendo que obtemos o coeficiente angular através da regressão linear a partir dos dados na janela de três meses em cada uma das linhas da base de dados respectiva.

A explicação da razão do emprego das variáveis em cada uma das situações de *feature engineering* são as consequentes:

- **variável 1**: identificar a proporção de aumento ou decréscimo no intervalo de janela utilizado (três meses),
- **variáveis 2 a 5**: averiguar se na janela de dados utilizados existe uma tendência de subida ou descida em relação aos dados do índice ou cotação específico,
- **variáveis 6 a 8**: verificar a taxa de variação do índice IBOV ou IFIX em relação aos demais dados de mercado, da análise AIB ou AII, respectivamente.

Durante as análises, testamos outro conjunto de variáveis obtidas com *feature engineering*, sendo que este conjunto foi composto por quatro variáveis. Porém, os resultados foram piores que os com a versão apresentada anteriormente. A partir destes dados é que são apresentados os resultados das tabelas dos Anexos A e B. Quando criamos as *feature engineering* utilizamos os dados do índice e dados de mercado não normalizados ainda.

6.2 Resultados Numéricos

Nas análises de tendência para ambos os índices, IBOV e IFIX, primeiramente adicionamos os dados do índice, em seguida os dados de mercado, ambos em janela de três meses e por último as variáveis que obtemos com *feature engineering*. Somente após realizamos a normalização dos dados, antes de aplicarmos estes nos modelos de Machine Learning.

Para a normalização dos dados utilizamos a função 'MinMaxScaler', da biblioteca scikit-learn da linguagem de programação Python, em que a normalização é dada pela equação (6.1) a seguir. Esta normalização 'MinMaxScaler' realiza o dimensionamento de cada variável explicativa individualmente para estar na faixa do conjunto de treinamento, por exemplo, entre 0 e 1 [67]. Desta maneira, fazendo com que os dados sejam comparáveis e não tendo discrepância devido à amplitude de qualquer uma das variáveis.

$$\begin{aligned} X_{i,std} &= \frac{X_i - X_{i,min}}{X_{i,max} - X_{i,min}} \\ X_{i,scaled} &= X_{i,std}(max - min) + min, \end{aligned} \tag{6.1}$$

sendo que $X_{i,max}$ e $X_{i,min}$ são os valores máximo e mínimo da coluna i da matriz de variáveis explicativas X e max e min é o máximo e mínimo que escolhemos, por padrão os valores de máximo e mínimo são 0 e 1.

Pela aplicação de Estatística básica nos dados observamos que ambos os índices têm classes desbalanceadas na variável resposta y. Por esse motivo, utilizamos no lugar da medida de avaliação acurácia as medidas de avaliação acurácia balanceada, precisão, recall e F1-score. Pois, caso utilizássemos a acurácia, os algoritmos utilizados pelos modelos de ML não diferenciariam a classe majoritária das demais, problema conhecido por Paradoxo da Acurácia [2]. Com a divisão dos dados de forma temporal com 20% para teste (validação) e 80% para treinamento, temos na Tabela 6.1 a proporção do desbalanceamento para ambos os índices do total de 245 meses de dados para o índice IBOV e 125 meses de dados para o índice IFIX. Nos dados consideramos na variável resposta y que quando o índice subiu do mês atual para o mês h à frente temos $y = 1$, e quando o índice desceu $y = 0$.

Tabela 6.1: Proporção do desbalanceamento das classes da variável resposta.

Índice	**h**	$y_{treinamento} = 1$	$y_{treinamento} = 0$	$y_{teste} = 1$	$y_{teste} = 0$
IBOV	1	0.58	0.42	0.63	0.37
	3	0.61	0.39	0.73	0.27
	6	0.61	0.39	0.77	0.23
	12	0.61	0.39	0.83	0.17
IFIX	1	0.67	0.33	0.64	0.36
	3	0.74	0.26	0.79	0.21
	6	0.72	0.28	0.75	0.25
	12	0.78	0.22	0.83	0.17

Fonte: O autor (2021).

Durante a análise de tendência realizamos análises com combinação de dados do índice, de mercado e de variáveis encontradas através de *feature engineering*, usando ou não cada uma destas bases de dados. Após essas análises, ainda, refazemos as análises utilizando x% das variáveis mais significativas, obtidas por meio da Análise de Variância (ANOVA), onde a partir deste método obtemos o valor F (*F Score*) que utilizamos para rankear as variáveis. Aplicando a ANOVA averiguamos se as variáveis possuem dependência de uma em relação às demais [61]. O objetivo aqui foi verificarmos se a retirada de alguma das variáveis sucederia em melhores resultados, em relação à medida de avaliação acurácia balanceada.

No Anexo C, temos as tabelas das variáveis mais significativas utilizando a ANOVA, separadas pelo índice da análise, AIB ou AII, e pelo horizonte de previsão h. Explorando essa análise, a importância de uma ou outra variável depende tanto do valor de h utilizado quanto dos índice, não havendo um resultado em comum. Desta forma, concluímos que em cada uma das previsões para h meses à frente temos um estudo diferente. A principal evidência que notamos é que a cotação do dólar e do ouro possuem bastante influência no índice IBOV na maioria dos casos e que o índice IBOV tem pouca influência no índice IFIX.

Os resultados da análise de tendência resultantes dos melhores modelos obtidos em cada caso são detalhados nas Seções 6.2.1 e 6.2.2. Para ver mais resultados, consultar as tabelas exibidas nos Anexos A e B.

Na realização das análises aplicando os modelos de Aprendizagem de Máquinas, para a obtenção dos melhores hiperparâmetros utilizamos um GridSearch, da biblioteca scikit-learn da linguagem de programação Python, passando para cada um dos modelos a lista de parâmetros subsequente

- Regressão Logística
  - solver: [‘lbfgs’],
  - penalty: [‘none’; ‘l2’],
  - class_weight: [‘None’; ‘balanced’],
  - C: $[10^{-3}; 10^{-2}; 10^{-1}; 0; 1; 10; 10^{2}]$, onde $\lambda = \dfrac{1}{2C}$,
- Floresta Aleatória
  - n_estimators: [10; 20; 50],
  - max_features: [‘auto’; ‘sqrt’],
  - class_weight: [‘None’; ‘balanced’],
- *XGBoost*
  - colsample_bytree: [0.1],
  - scale_pos_weight: [$y_{train}.value\_counts()[0]$ / $y_{train}.value\_counts()[1]$], quantidade das instâncias negativas dividido pela quantidade das instâncias positivas,
- Máquina de Vetores Suporte
  - C: $[10^{-7}; 10^{-5}; 10^{-3}; 1]$,
  - kernel: [‘rbf’],
  - gamma: [‘auto’; ‘scale’], sendo que ‘auto’ = 1 / (número de variáveis) e ‘scale’ = 1 / (número de variáveis $\cdot$ variância de X),
  - class_weight: [‘None’; ‘balanced’],
- Rede Neural
  - epochs: [150],
  - batch_size: [40],
  - learning_rate: [0.05; 0.1],
  - optimizer: [’sgd’; ’Adam’],
  - activation: [’tanh’].

sendo a semente = 6 (*random state*) utilizado durante nas análises. Os outros parâmetros dos modelos ou apresentaram resultados piores, ou não fizeram diferença na modelagem ou apresentaram um tempo computacional alto em comparação com os outros modelos testados. Nestas situações, ficamos com o valor padrão (*default*) que pode ser consultado na documentação do respectivo modelo no site do pacote Scikit-learn do Python. Durante a explicação dos melhores modelos conseguidos, citaremos quais foram os hiperparâmetros obtidos.

6.2.1 Análise dos dados do índice Ibovespa

		Valor Predito	
		Descida (0)	Subida (1)
Valor Real	Descida (0)	72.22% 13	27.78% 5
	Subida (1)	35.48% 11	64.52% 20

Figura 6.1: Matriz de confusão do modelo com dados do índice IBOV, sem dados de mercado e sem *feature engineering*.
Fonte: O autor (2021).

Na Figura 6.1, temos a matriz de confusão do melhor modelo, segundo a acurácia balanceada, somente com uso dos dados do índice IBOV com a janela de dados do índice de três meses como mostrado anteriormente. O melhor modelo que obtemos foi a RL com previsão de um mês à frente ($h = 1$). A partir desse resultado, obtemos os valores das medidas de avaliação

- Acurácia: 67.35%,
- Acurácia Balanceada: 68.37%,
- Precisão: 0.80,
- Recall: 0.65,
- F1-Score: 0.71,

e os melhores hiperparâmetros usados por esse modelo foram

- solver: [‘lbfgs’],
- penalty: [‘none’],
- class_weight: [‘balanced’],
- C: [10^{-3}].

		Valor Predito	
		Descida (0)	Subida (1)
Valor Real	Descida (0)	100.00% 8	0.00% 0
	Subida (1)	41.03% 16	58.97% 23

Figura 6.2: Matriz de confusão do modelo com dados do índice IBOV, dados de mercado e sem *feature engineering* aplicando ANOVA.
Fonte: O autor (2021).

Aplicando os modelos de ML aos dados encontramos que o melhor modelo com acréscimo dos dados de mercado foi o modelo SVM com previsão de um ano à frente ($h = 12$), este resultado teve a aplicação da ANOVA para ficarmos com 90% das variáveis mais significativas, pois apresentou melhor resultado que o modelo usando todos os dados do índice e de mercado. Na Figura 6.2 temos a matriz de confusão deste modelo e as medidas de avaliação são as seguintes

- Acurácia: 65.96%,
- Acurácia Balanceada: 79.49%,
- Precisão: 1.00,
- Recall: 0.59,
- F1-Score: 0.74,

e os melhores hiperparâmetros encontrados usando GridSearch foram

- C: [1],
- kernel: ['rbf'],
- gamma: ['auto'],
- class_weight: ['balanced'].

		Valor Predito	
		Descida (0)	**Subida (1)**
Valor Real	**Descida (0)**	100.00% 8	0.00% 0
	Subida (1)	56.41% 22	43.59% 17

Figura 6.3: Matriz de confusão do modelo com dados do índice IBOV, dados de mercado e *feature engineering*.
Fonte: O autor (2021).

Neste passo também temos o emprego de variáveis obtidas por *feature engineering*, o melhor modelo encontrado neste caso foi o modelo XGB com previsão de um ano à frente ($h = 12$), cuja matriz de confusão é mostrada na Figura 6.3. As medidas de avaliação desse modelo foram

- Acurácia: 53.19%,
- Acurácia Balanceada: 71.79%,
- Precisão: 1.00,
- Recall: 0.44,

- F1-Score: 0.61,

e os melhores hiperparâmetros encontrados para esse modelo são os seguintes

- colsample_bytree: [0.1],
- scale_pos_weight: [0.6429].

		Valor Predito	
		Descida (0)	**Subida (1)**
Valor Real	**Descida (0)**	100.00% 8	0.00% 0
	Subida (1)	43.59% 17	56.41% 22

Figura 6.4: Matriz de confusão do modelo com dados do índice IBOV, dados de mercado, *feature engineering* e usando a ANOVA.
Fonte: O autor (2021).

Por fim, aplicamos a ANOVA para seleção de 90% das variáveis mais significativas, cuja porcentagem foi que resultou no melhor modelo. Vemos nesta análise, que o melhor modelo foi o SVM com previsão de ano à frente ($h = 12$). A matriz de confusão é mostrada na Figura 6.4, em que as medidas de avaliação encontradas foram

- Acurácia: 63.83%,
- Acurácia Balanceada: 78.21%,
- Precisão: 1.00,
- Recall: 0.56,
- F1-Score: 0.72,

e os melhores hiperparâmetros usados neste modelo são

- C: [1],
- kernel: [‘rbf’],
- gamma: [‘auto’],
- class_weight: [‘balanced’],
- probability: [False].

Em resumo, na análise com o índice IBOV, verificamos que os melhores resultados em relação à acurácia balanceada foi com o modelo SVM com $h = 12$ com 79.49%, quando utilizamos dados do índice e dados de mercado, além de selecionarmos as 90% das variáveis mais significativas. Mostrando que no período analisado quando adicionamos mais variáveis explicativas, obtemos melhores resultados, porém somente não tendo efeito a utilização de variáveis obtidas por *feature engineering*. Decidimos considerar o período

de Janeiro de 2001 a Maio de 2021 para que as análises com dados de mercado e *feature engineering* ficassem padronizadas porque não tínhamos dados de mercado em algumas variáveis antes do ano de 2001.

Com relação à medida de avaliação F1-Score, entre os modelos com melhor acurácia balanceada, percebemos que o melhor modelo se manteve o modelo SVM com $h = 12$, com F1-Score de 0.74, no cenário com utilização de dados do índice, de mercado e 90% das variáveis mais significativas obtidas pela ANOVA. Como as conclusões se mantiveram, podemos usar tanto acurácia balanceada quanto F1-Score. Porém, optamos por utilizar os resultados da acurácia balanceada como principal medida de avaliação.

6.2.2 Análise dos dados do índice de Fundos de Investimentos Imobiliários

		Valor Predito	
		Descida (0)	**Subida (1)**
Valor Real	**Descida (0)**	44.44% 4	55.56% 5
	Subida (1)	18.75% 3	81.25% 13

Figura 6.5: Matriz de confusão do modelo com dados do índice IFIX, sem dados de mercado e sem *feature engineering*.
Fonte: O autor (2021).

Usando somente dados do índice, obtemos que o melhor modelo foi o RL com previsão de um mês à frente ($h = 1$), na qual temos sua matriz de confusão na Figura 6.5. A partir da matriz de confusão temos as seguintes medidas de avaliação

- Acurácia: 68.00%,
- Acurácia Balanceada: 62.85%,
- Precisão: 0.72,
- Recall: 0.81,
- F1-Score: 0.76,

e os melhores hiperparâmetros obtidos com GridSearch para esse modelo foram

- solver: [‘lbfgs’],
- penalty: [‘none’],
- class_weight: [‘None’],
- C: $[10^{-3}]$.

Na Figura 6.6, mostramos o resultado do modelo RN com previsão de um mês à frente ($h = 1$) que foi o melhor modelo com acréscimo de dados de mercado, cujas medidas de avaliação encontradas são

		Valor Predito	
		Descida (0)	Subida (1)
Valor Real	Descida (0)	66.67% 6	33.33% 3
	Subida (1)	37.50% 6	62.50% 10

Figura 6.6: Matriz de confusão do modelo com dados do índice IFIX, dados de mercado e sem *feature engineering*.
Fonte: O autor (2021).

- Acurácia: 64.00%,
- Acurácia Balanceada: 64.58%,
- Precisão: 0.77,
- Recall: 0.62,
- F1-Score: 0.69,

e os melhores hiperparâmetros foram

- epochs: [150],
- batch_size: [40],
- learning_rate: [0.05],
- optimizer: ['Adam'],
- activation: ['tanh'].

		Valor Predito	
		Descida (0)	Subida (1)
Valor Real	Descida (0)	55.56% 5	44.44% 4
	Subida (1)	43.75% 7	56.25% 9

Figura 6.7: Matriz de confusão do modelo com dados do índice IFIX, dados de mercado e com *feature engineering*.
Fonte: O autor (2021).

O melhor modelo utilizando dados de mercado e *feature engineering*, que encontramos foi o modelo RL com previsão de um mês à frente ($h = 1$), que tem sua matriz de confusão apresentada na Figura 6.7, cujas medidas de avaliação são as seguintes

- Acurácia: 56.00%,
- Acurácia Balanceada: 55.90%,

- Precisão: 0.69,
- Recall: 0.56,
- F1-Score: 0.62,

e os melhores hiperparâmetros encontrados para esse modelo foram

- solver: ['lbfgs'],
- penalty: ['l2'],
- class_weight: ['balanced'],
- C: [10].

		Valor Predito	
		Descida (0)	**Subida (1)**
Valor Real	**Descida (0)**	50.00% 2	50.00% 2
	Subida (1)	0.00% 0	100.00% 19

Figura 6.8: Matriz de confusão do modelo com dados do índice IFIX, dados de mercado, *feature engineering* e usando a ANOVA.
Fonte: O autor (2021).

Por último, usando, além dos anteriores, a ANOVA para encontrar as 95% das variáveis mais significativas, encontramos que o melhor modelo foi o XGB com previsão de um ano à frente ($h = 12$), na Figura 6.8. As medidas de avaliação obtidas com esse modelo foram as seguintes

- Acurácia: 91.30%,
- Acurácia Balanceada: 75.00%,
- Precisão: 0.90,
- Recall: 1.00,
- F1-Score: 0.95,

e os melhores hiperparâmetros obtidos com GridSearch para este modelo foram os seguintes

- colsample_bytree: [0.1],
- scale_pos_weight: [0.2754].

Em relação ao índice IFIX, explorando os resultados percebemos que o melhor modelo em relação à acurácia balanceada foi o modelo XGB com $h = 12$ e uso de variáveis de mercado e *feature engineering* utilizando 95% das variáveis mais significativas obtidas pela ANOVA, com 75.00%. Verificamos que o uso da ANOVA com a retirada de uma das variáveis apresentou melhora nos resultados, sendo o melhor modelo para a análise com o índice IFIX. Isso mostra que a adição de variáveis do mercado e *feature engineering* com retirada de variáveis a partir da ANOVA, ajudaram a melhorar o modelo aumentando o valor da medida de avaliação acurácia balanceada, sendo que a melhora foi de 12.15% em relação modelo somente com dados do índice e de 10.42% em relação ao modelo usando dados de mercado, que representou grande valor quando analisamos a quantidade monetária que se pode ganhar. Além disso, com a adição de dados de mercado a acurácia balanceada melhorou e quando adicionamos mais variáveis obtidas por *feature engineering* o modelo piorou, somente melhorando para o caso específico com $h = 12$ usando o modelo XGB quando removemos a variável menos significativa, que vemos na Tabela C.4 do Anexo C.

Agora, analisando em relação à medida de avaliação F1-Score, entre os modelos com melhor acurácia balanceada, percebemos que o melhor modelo que obtemos se manteve sendo o XGB com $h = 12$, onde utilizamos 95% de todas as variáveis, índice, mercado e *feature engineering*, com F1-Score de 0.95. Em segundo lugar temos o modelo RL com a utilização somente dos dados do índice, com F1-Score de 0.76. O que mostra um resultado semelhante em relação à medida de avaliação acurácia balanceada, quando analisado o melhor modelo entre as as opções verificadas. Neste caso podemos usar qualquer uma das medidas de avaliação, mas fixamos a acurácia balanceada como medida de avaliação principal.

Os resultados encontrados são de difícil aplicação na vida real, pois dependem tanto do período dos dados em que se esteja utilizando, quando do valor de h, que é a quantidade de meses à frente que estamos prevendo. Cada vez que mudamos o período, obtemos um resultado diferente e consequentemente o melhor modelo tem alteração em cada uma destas situações. Além disso, a mudança nos resultados depende também dos dados de mercado e das variáveis obtidas com *feature engineering* utilizadas, sendo o melhor resultado aplicando algumas dessas variáveis explicativas em uma situação e outras em alguma situação diferente. Desta forma, as conclusões obtidas são referentes ao cenário que utilizamos, ou seja, para os dados de mercado e *feature engineering* aplicados e também para o período específico dos dados considerados.

Capítulo 7

Comparação de Fundos Imobiliários com Outros Índices

Neste capítulo, fazemos a comparação entre o índice IFIX com os índices IBOV, IDIX, SMLL e IMOB. Reproduzimos a análise realizada em [69] modificando o período de análise dos dados para verificarmos se as mudanças na composição das carteiras do Ibovespa e do IFIX trazem ou não variações significativas em relação aos resultados apresentados no artigo. Além disso, essa análise comparativa é realizada com o intuito de verificarmos o desempenho do IFIX em relação a outros índices financeiros que possuem relação com fundos imobiliários e ações.

7.1 Resultados Numéricos

Nesta seção apresentamos os resultados das análises e comparações entre os índices IFIX, IBOV, IDIV, SMLL e IMOB. Inicialmente, realizamos uma análise descritiva dos dados. Em seguida, fazemos algumas análises usando os testes estatísticos de normalidade, igualdade de média e homogeneidade de variância, descritos no Capítulo 5. Esta comparação foi baseada no artigo "Comparação do risco-retorno do IFIX com IBOVESPA, IDIV, SMLL e IMOB" [69]. Durante essa análise utilizamos os dados da forma log-retorno, conforme explicado no Capítulo 5.

As análises foram realizadas considerando as comparações dos índices financeiros usando dados:

1. de todos os meses da base de dados;
2. dos meses em que o índice IBOV está em período de alta;
3. dos meses em que o índice IBOV está em período de baixa;
4. dos meses em que ambos os índices IFIX e IBOV estão em alta;
5. dos meses em que ambos os índices IFIX e IBOV estão em baixa;

sendo que os meses em alta são os meses em que o log-retorno do índice em questão é maior ou igual a zero e os meses em baixa quando o log-retorno do índice está menor que zero.

7.1.1 Análise Descritiva

Os resultados da análise descritiva podem ser vistos na Tabela 7.1. Nesta tabela, apresentamos os valores dos retornos (médias) e dos riscos (desvio-padrão) dos índices considerados com o intuito de fazermos uma comparação. Podemos perceber que o índice que obteve maior retorno com menor risco foi o IFIX, seguindo do IDIV que possui risco perto do dobro do IFIX. Isso se deve ao fato de o índice IFIX ter menor volatilidade. Como no cenário mundial passamos por crises, verificamos isso refletido nos investimentos, e averiguamos que o índice IFIX se recupera mais rápido nestes períodos, quando comparado com outros índices. Também notamos, na análise descritiva, que o índice IMOB está com retorno nulo, mostrando um processo mais demorado de recuperação ao melhor momento anterior, verificando uma maior volatilidade. Estes resultados descritivos são mostrados na Figura 7.1.

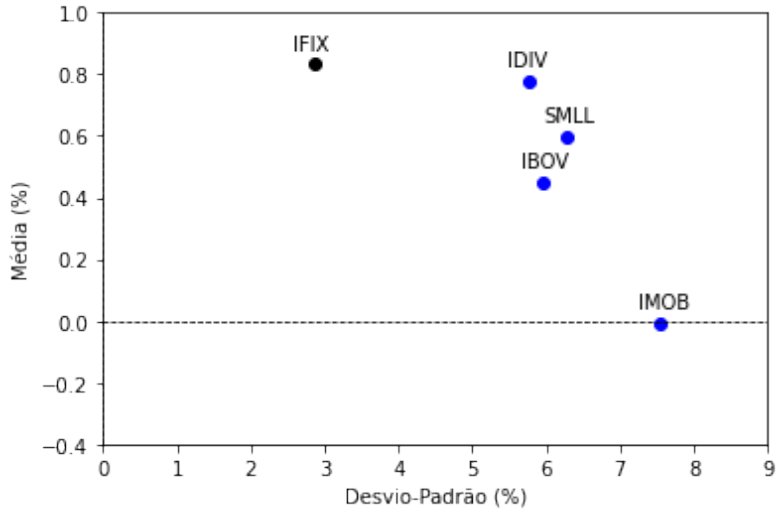

Figura 7.1: Desvio-padrão e média dos índices comparados.
Fonte: O autor (2020).

Tabela 7.1: Desvio-padrão, média e correlação entre os índices.

Índice	**Média**	**Desvio Padrão**	**IFIX**	**IBOV**	**IDIV**	**SMLL**	**IMOB**
IFIX	0.832%	2.860%	1.000	-	-	-	-
IBOV	0.448%	5.969%	0.651	1.000	-	-	-
IDIV	0.777%	5.761%	0.695	0.911	1.000	-	-
SMLL	0.595%	6.277%	0.705	0.908	0.879	1.000	-
IMOB	-0.007%	7.533%	0.632	0.856	0.829	0.925	1.000

Fonte: O autor (2021).

Nas análises, com todos os 125 meses de dados, a partir da análise de correlação, averiguamos que o índice IFIX possui correlação forte (entre 0.70 e 0.89) com os índices IDIV e SMLL e correlação moderada (entre 0.40 e 0.69) com os índices IBOV e IMOB, sendo que a maior correlação do IFIX foi de 0.705 com o índice SMLL e a menor correlação foi de 0.632 com o índice IMOB. Neste aspecto, os resultados se mantiveram com correlação positiva, apesar dos valores passarem a ter uma correlação um pouco maior a

medida que o período de dados aumenta. Este resultado não era esperado, pois o índice imobiliário (IMOB) trata-se do mesmo macrossetor, segundo [69]. Além disso, o índice IBOV apresentou uma correlação alta com os índices, com exceção do IFIX. Com as análises, fica evidente que o índice IFIX é um pouco diferente dos outros índices, possuindo características de volatilidade e recuperação melhor em caso de queda ocorrida por crises na economia.

7.1.2 Análise de todo o período

Na exploração e comparação dos cinco índices com dados de todo o período, usamos o teste de normalidade Kolmogorov-Smirnov (KS) para sabermos se os dados seguem a distribuição normal e intuirmos se usamos testes paramétricos ou testes não paramétricos para realizarmos comparação da média e variância dos índices. Segundo os resultados obtidos na Tabela 7.2, observamos que não existem indícios para rejeitarmos a hipótese nula H_0, pois os valores p-valor foram superiores a 0.100, isto é, a probabilidade de evidência contra a hipótese nula é pequena o que implica que o p-valor está na região de aceitação da hipótese H_0, ou ainda, por que o valor da estatística do teste KS é menor que o valor crítico ($D_{cal} < D_{crit}$), ver Figura 7.2. Consequentemente os dados seguem a distribuição normal e podemos usar os testes paramétricos em todas as comparações.

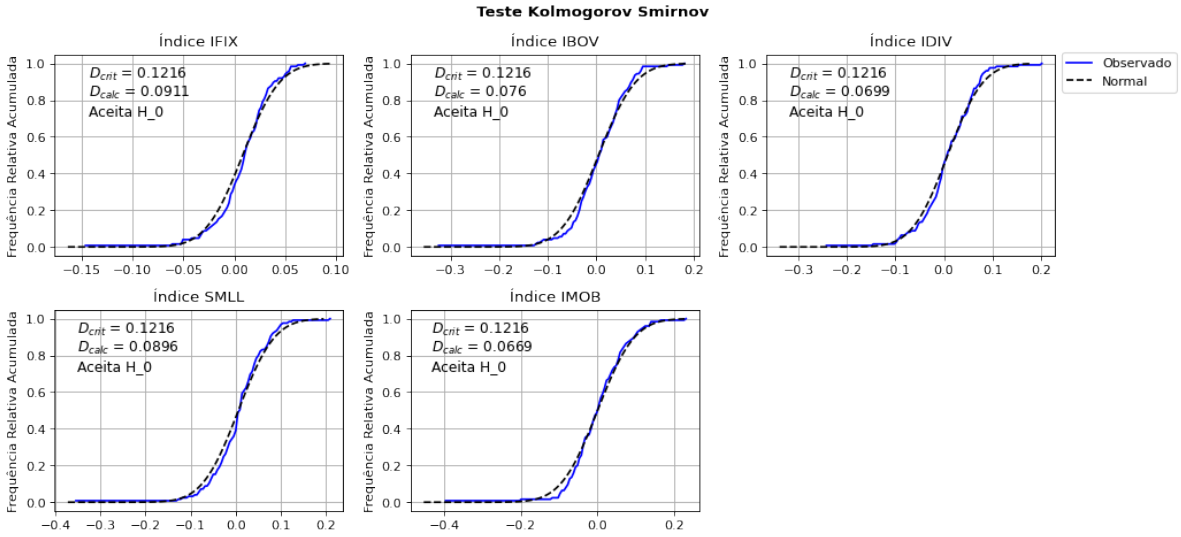

Figura 7.2: Gráficos dos resultados do teste de hipótese Kolmogorov-Smirnov para averiguar a suposição de normalidade considerando todo o período.
Fonte: O autor (2021).

Na investigação em busca de evidências de que os índices têm comportamento similar, utilizando o teste t bicaudal com dados pareado para comparação das médias dos índices, Tabela 7.2, e concluímos que as diferenças das médias são estatisticamente significantes ao retorno do IFIX (0.00821), porque $-t_{\alpha/2} = -1.979 < t_{cal} < 1.979 = t_{\alpha/2}$ com nível de significância de $\alpha = 0.05$ conforme mostrado na Figura 7.3. Desta forma, as médias dos índices são significantemente iguais à média do índice IFIX.

Também realizamos o Teste F para verificar a homogeneidade das variâncias dos índices e verificar se os índices possuem dispersão semelhante em relação à média. Com

a aplicação do Teste F, acabamos rejeitando a hipótese H_0, pois os valores de p-valor encontrados em cada comparação foram nulos, ver Tabela 7.2, o que significa que o p-valor está na região de rejeição da hipótese H_0, isto é, $F_{calc} < F_{\alpha/2}$ ou $F_{calc} > F_{1-\alpha/2}$ com nível de significância de $\alpha = 0.05$ conforme Figura 7.4. Logo, a probabilidade de evidência contra a hipótese H_0 é grande. Portanto, o índice IFIX possui volatilidade menor que os outros índices comparados, além de um risco (desvio-padrão) menor, como podemos notar pela análise descritiva. Nas análises de comparação da variância utilizamos os dados log-retorno na função do teste de F da linguagem de programação Python, isto é, não passamos na programação os dados já na forma pareada.

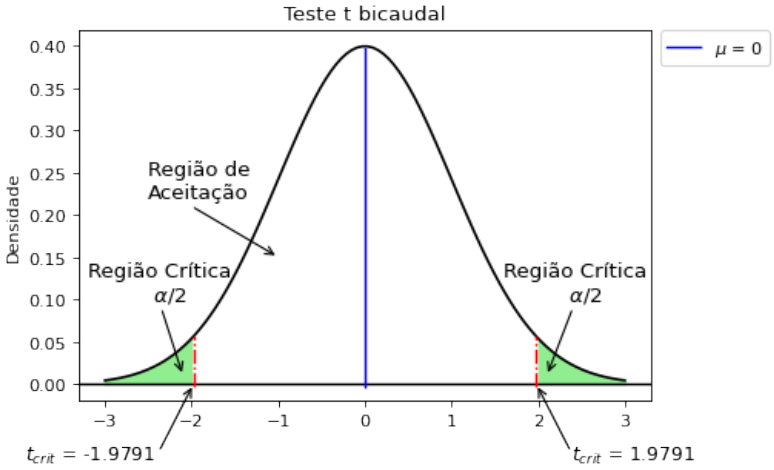

Figura 7.3: Gráfico do teste de hipótese t de Student para comparação da média considerando todo o período.
Fonte: O autor (2021).

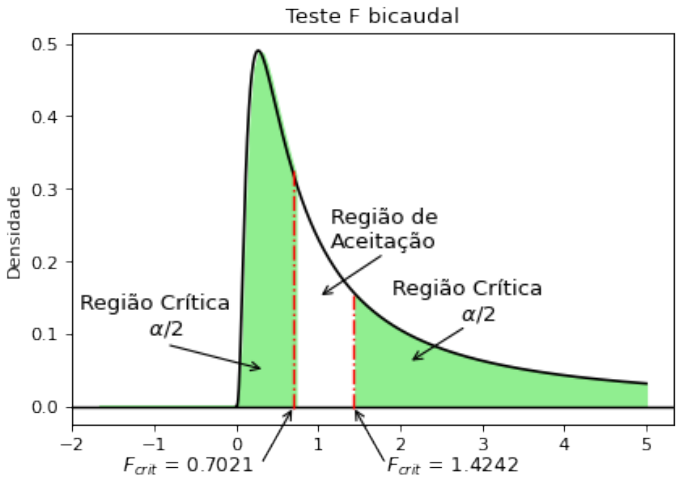

Figura 7.4: Gráfico do teste de hipótese F para comparação da variância considerando todo o período.
Fonte: O autor (2021).

Tabela 7.2: Média, desvio-padrão, teste de normalidade, teste de igualdade de média e teste de igualdade de variância para a análise com dados de todo o período.

Estatística ou Teste	**IFIX**	**IBOV**	**IDIV**	**SMLL**	**IMOB**
Média	0.832%	0.448%	0.777%	0.595%	-0.007%
Desvio-Padrão	2.860%	5.969%	5.761%	6.277%	7.533%
Normalidade ($p-valor$)	0.237	0.448	0.564	0.253	0.625
Normalidade (D_{calc})	0.091	0.076	0.070	0.090	0.067
Igualdade de Média ($p-valor$)	-	0.359	0.886	0.577	0.131
Igualdade de Média (t_{calc})	-	0.920	0.144	0.560	1.521
Igualdade de Variância ($p-valor$)	-	0.000	0.000	0.000	0.000
Igualdade de Variância (F_{calc})	-	0.230	0.247	0.208	0.144

Fonte: O autor (2021).

7.1.3 Análises do IBOV em período de alta e em período de baixa

Com a análise descritiva, podemos perceber que o retorno (média) no período foi pequeno comparado com o risco (desvio-padrão), isso mostra não compensa realizar investimentos nestes índices. Deste modo, como no artigo original [69], para uma análise mais completa e detalhada, repetimos os estudos da parte com todos os dados dos 125 meses, dividindo a análise em duas partes, a primeira considerando os meses com o índice IBOV em período de alta e na outra com os meses em que o índice IBOV esteve em período de baixa durante o período dos dados. Os meses em alta são os meses em que o log-retorno do índice em questão é maior ou igual a zero e os meses em baixa quando o log-retorno do índice é menor que zero. Pela análise exploratória, verificamos que há 68 meses em que o índice IBOV estava em alta (54.4% dos dados) e 57 meses em baixa (45.6% dos dados).

Os resultados da análise do índice IBOV em período de alta e em período de baixa podem ser vistos nas Tabelas 7.3 e 7.4, respectivamente. Destas tabelas, podemos apurar que o desempenho do IBOV se manteve, isto é, o risco permaneceu grande com um retorno pequeno. Também percebemos que os retornos dos índices IFIX, IDIV e SMLL tiveram um pouco de aumento, porém com o risco dos índices também aumentando de forma proporcional nos meses de alta, e nos meses de baixa ficando semelhantes com exceção do índice IFIX. Além disso, o índice IFIX continuou tendo um retorno maior que os outros índices.

Aplicando o teste de normalidade nos meses em alta, verificamos que o pressuposto de normalidade somente é aceito para o índice IFIX, visto que, conforme os resultados apresentados na Tabela 7.3, rejeitamos a hipótese nula, isto é, de acordo com a Figura 7.5 temos que $D_{cal} < D_{crit} = 0.1649$, encontrado na tabela do teste KS. Para os demais índices a probabilidade de evidência contra a hipótese nula é grande. Portanto, para essa parte da análise com dados onde o índice IBOV estava em alta, somente aplicamos testes não paramétricos para comparação da média e variância.

Para comparação do retorno, efetuamos o uso do teste Wilcoxon para dados pareados. Os resultados são apresentados na Tabela 7.3 e obtemos que $-1.996 = -z_{\alpha/2} < W_{calc} = z_{calc} < z_{\alpha/2} = 1.996$, com nível de significância $\alpha = 0.05$, ver Figura 7.6. Portanto,

concluímos que somente para o índice IDIV o teste decorre da rejeição da hipótese H_0, ou seja, a não rejeição da hipótese que os retornos são estatisticamente iguais. Doravante, diferentemente aos dados completos com IBOV em que todos os retornos são estatisticamente significativos, neste caso do índice IBOV com período em alta o resultado foi diferente, mostrando que se realizarmos uma aplicação no índice IDIV provavelmente teremos um retorno diferente ou não similar aos encontrados com aplicações em algum dos demais índices comparados.

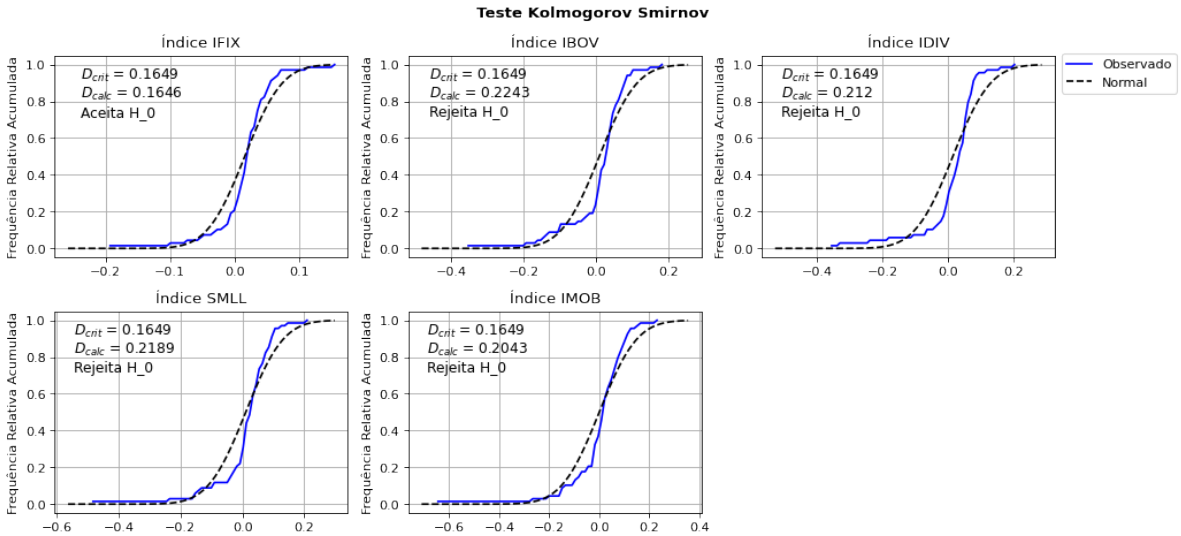

Figura 7.5: Gráficos dos resultados do teste de hipótese Kolmogorov-Smirnov para averiguar a suposição de normalidade considerando o período com IBOV em alta.
Fonte: O autor (2021).

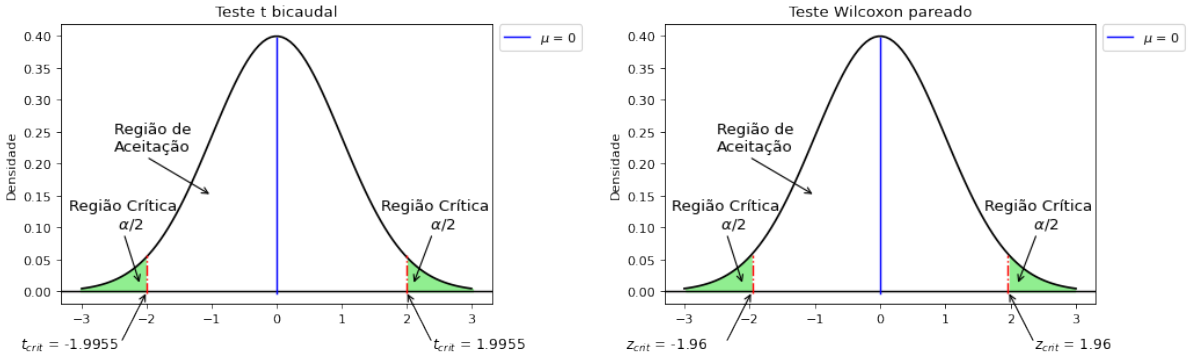

Figura 7.6: Gráficos dos testes de hipóteses t e Wilcoxon para comparação da média considerando período com IBOV em alta.
Fonte: O autor (2021).

Em seguida, para fazermos a comparação de homogeneidade de variância, aplicamos o teste não paramétrico Levene. Para todos os índices, obtivemos que o risco (desvio-padrão) não é significantemente igual ao risco do índice IFIX. Pois, nos testes realizados sempre rejeitamos a hipótese nula H_0, em que temos a hipótese de igualdade da variância entre o índice em questão e o índice IFIX, ou seja, ou $F_{\alpha/2} = 0.617 < L_{calc} < F_{1-\alpha/2} = 1.621$ de acordo com Figura 7.7. Assim sendo, como na análise com dados completos

encontramos que o índice IFIX tem menor volatilidade mesmo somente nos meses com o índice IBOV em alta.

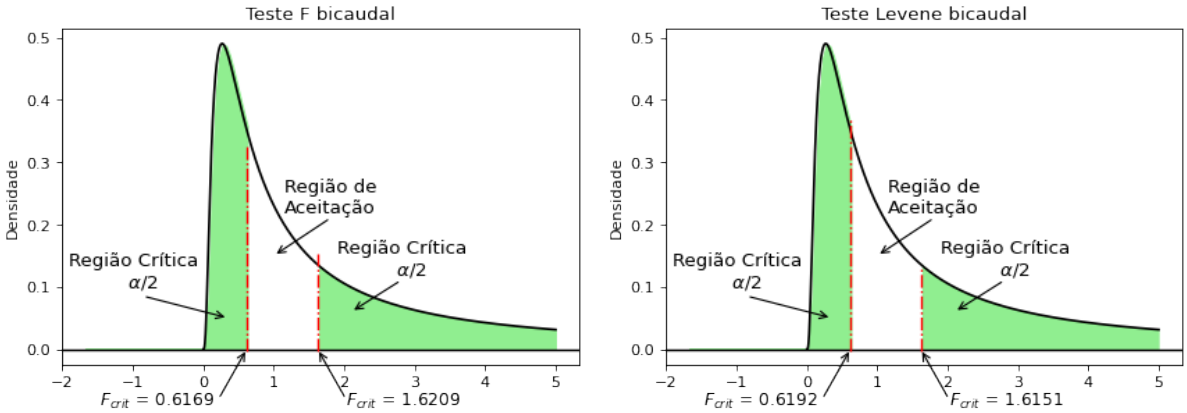

Figura 7.7: Gráficos dos testes de hipóteses F e Levene para comparação da variância considerando período com IBOV em alta.
Fonte: O autor (2021).

Tabela 7.3: Média, desvio-padrão, teste de normalidade, teste de igualdade de média e teste de igualdade de variância para os meses de alta do índice IBOV.

Estatística ou Teste	**IFIX**	**IBOV**	**IDIV**	**SMLL**	**IMOB**
Média	1.529%	0.824%	1.428%	1.093%	-0.011%
Desvio-Padrão	4.519%	8.118%	8.968%	9.506%	11.753%
Normalidade $(p-valor)$	0.045	0.002	0.004	0.002	0.006
Normalidade (D_{calc})	0.165	0.224	0.212	0.219	0.204
Igualdade de Média $(p-valor)$	-	0.743	0.060	0.344	0.536
Igualdade de Média (t_{calc}/W_{calc})	-	-0.529	-2.047	-1.134	-0.813
Igualdade de Variância $(p-valor)$	-	0.005	0.005	0.003	0.002
Igualdade de Variância (F_{calc}/L_{calc})	-	8.057	8.156	8.722	14.422

Fonte: O autor (2021).

Agora, nesse passo, realizamos a análise para o período em que o índice IBOV está em baixa. Realizando o teste de normalidade Kolmogorov Smirnov, verificamos que os índices IFIX e SMLL possuem normalidade nos dados, ou seja, não rejeitamos a hipótese nula H_0 no teste KS, pois $D_{cal} < D_{crit} = 0.1801$, conforme resultados da Tabela 7.4 e Figura 7.8. Então, para o índice SMLL usamos testes paramétricos para comparação da média e da variância com o índice IFIX, e testes não paramétricos para os demais índices.

Comparando o retorno do índice IFIX com os outros índices, usamos o teste t de Student para os índices SMLL, e teste Wilcoxon para o outros índices. Desta maneira, temos as seguintes desigualdades dos métodos $-2.003 = -t_{\alpha/2} < t_{calc} < t_{\alpha/2} = 2.003$ e $-1.960 = -z_{\alpha/2} < W_{calc} = z_{calc} < z_{\alpha/2} = 1.960$ respectivamente, com nível de significância $\alpha = 0.05$, ver Figura 7.9. Logo, com os resultados da Tabela 7.4 concluímos que somente para o índice IMOB averiguamos que não possui retorno estatisticamente significante em comparação com o retorno do IFIX (0.01821) com dados do índice IBOV em baixa.

Ou seja, nos testes só rejeitamos a hipótese nula H_0 com o teste Wilcoxon usado em comparação com o índice IMOB. Assim, concluímos que existe somente diferença em investir no índice IMOB em comparação com o índice IFIX.

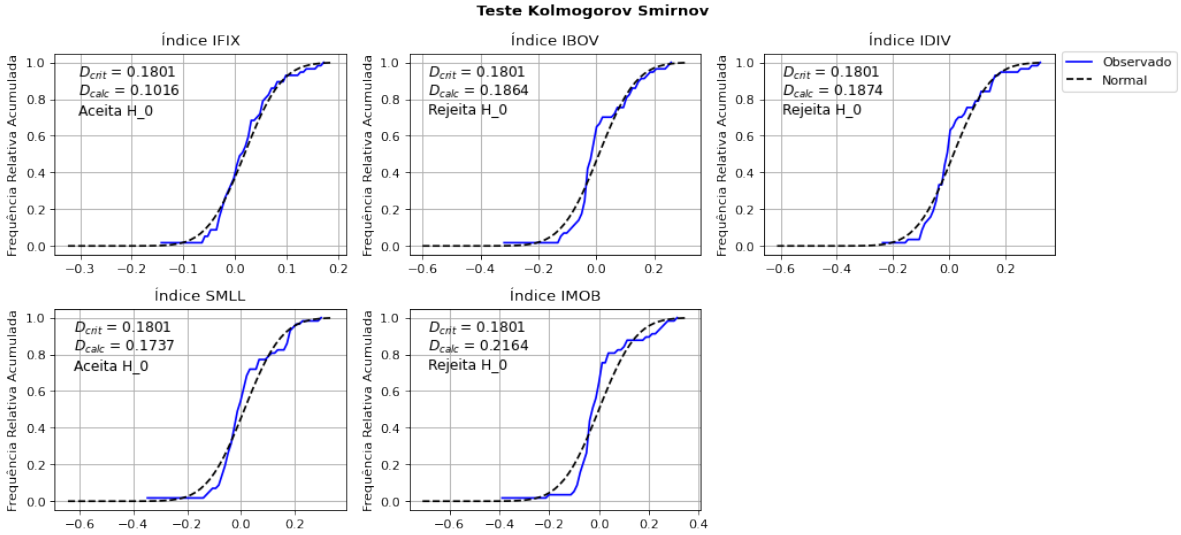

Figura 7.8: Gráficos dos resultados do teste de hipótese Kolmogorov-Smirnov para averiguar a suposição de normalidade considerando o período com IBOV em baixa.
Fonte: O autor (2021).

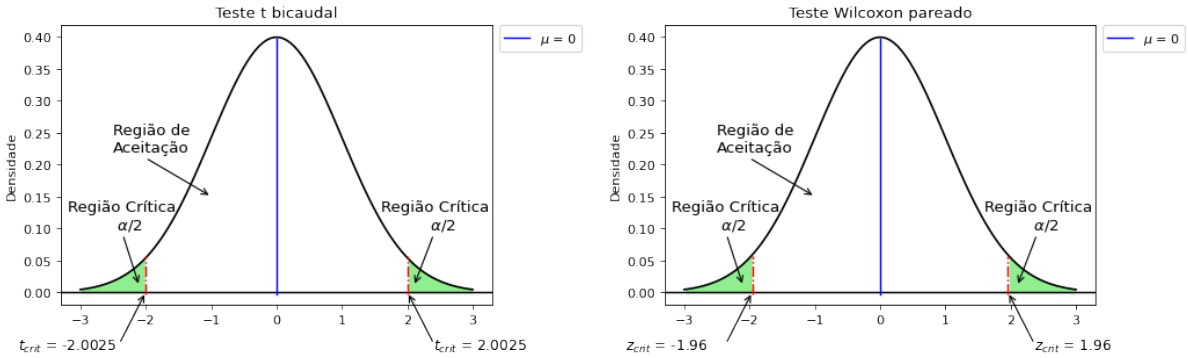

Figura 7.9: Gráficos dos testes de hipóteses t e Wilcoxon para comparação da média considerando período com IBOV em baixa.
Fonte: O autor (2021).

Por último, realizando os testes para verificar a homogeneidade entre os índices quando o índice IBOV está em baixa, encontramos o resultado aplicando o teste F para o índice SMLL e o teste Levene para os demais índices e rejeitamos a hipótese nula H_0 para todos os índices, concluindo que os índices com dados referentes ao período do IBOV em baixa também não possuem variância igual, ou seja, o risco atrelado aos índices é diferente. Chegamos a essa decisão pelos resultados da Tabela 7.4 em conjunto com a Figura 7.10.

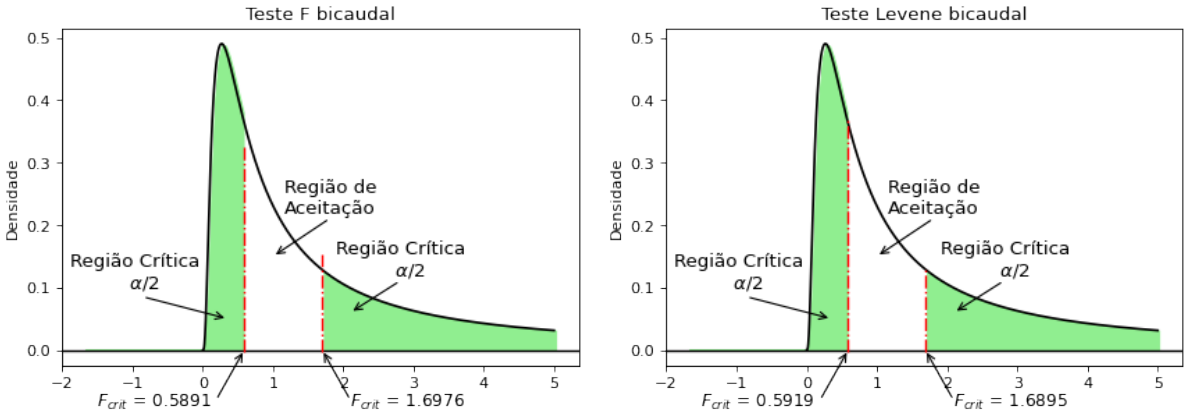

Figura 7.10: Gráficos dos testes de hipóteses F e Levene para comparação da variância considerando período com IBOV em baixa.
Fonte: O autor (2021).

Tabela 7.4: Média, desvio-padrão, teste de normalidade, teste de igualdade de média e teste de igualdade de variância para os meses de baixa do índice IBOV.

Estatística ou Teste	**IFIX**	**IBOV**	**IDIV**	**SMLL**	**IMOB**
Média	1.821%	0.943%	1.638%	1.271%	0.023%
Desvio-Padrão	5.649%	10.071%	10.394%	10.818%	11.709%
Normalidade $(p - valor)$	0.558	0.033	0.032	0.057	0.008
Normalidade (D_{calc})	0.102	0.186	0.187	0.174	0.216
Igualdade de Média $(p - valor)$	-	0.099	0.412	0.545	0.007
Igualdade de Média (t_{calc}/W_{calc})	-	-1.831	-1.025	0.609	-2.864
Igualdade de Variância $(p - valor)$	-	0.001	0.001	0.000	0.005
Igualdade de Variância (F_{calc}/L_{calc})	-	11.248	12.293	0.273	8.093

Fonte: O autor (2021).

Assim, com a separação dos dados em período do índice IBOV em alta e em baixa, não encontramos evidências para afirmar que a volatilidade dos índices acontece da mesma maneira, o que confirma a suspeita que o índice IFIX tem uma volatilidade menor que os demais índices comparados.

7.1.4 Análises do IBOV e IFIX em período de alta e em período de baixa

Pelo resultado obtido no estudo da seção anterior, pensamos em uma última análise separando os dados nos meses em que tanto o índice IFIX quanto o índice IBOV estão em alta e nos meses em que ambos os índices estão em baixa. Para tal, realizamos a comparação somente entre os dois índices: IFIX e IBOV. No artigo em que nos baseamos [69], os autores denominam essa estudo como análise de robustez, que possui o intuito de verificarmos se os resultados da análise de comparação da variância entre os índices apresentam resultados diferentes ou se mantêm parecidos, porque notamos que nos dados existem meses que os índices IBOV e IFIX subiam e que outros caíam simultaneamente. Durante esta análise, como na anterior usamos os mesmos testes estatísticos para comparação da média

e variância dos índices. Em uma análise exploratória identificamos que existem 54 meses com os índices IFIX e IBOV em alta ao mesmo tempo (43.2% dos dados) e 27 meses com ambos os índices em baixa simultaneamente (21.6% dos dados).

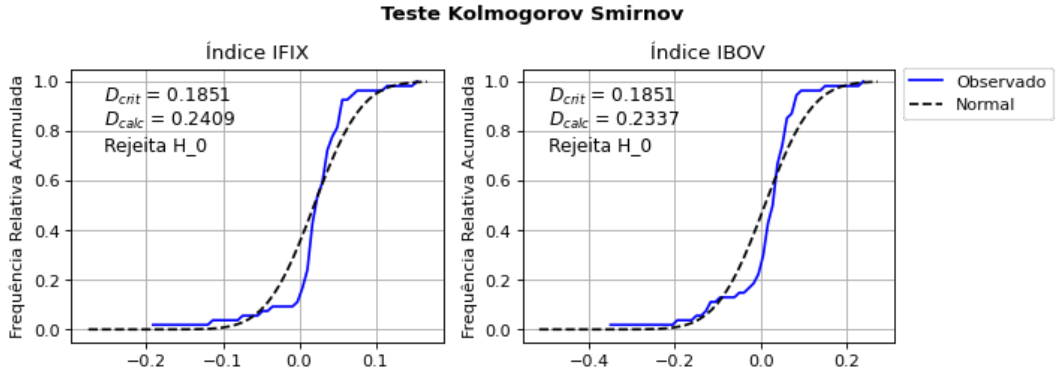

Figura 7.11: Gráficos dos resultados do teste de hipótese Kolmogorov-Smirnov para averiguar a suposição de normalidade considerando o período com IFIX e IBOV em alta.
Fonte: O autor (2021).

Para o caso em que ambos os índices IBOV e IFIX estão em período de alta, obtivemos a não aceitação da hipótese de normalidade do teste KS, tanto para o índice IFIX quanto para o índice IBOV, ver Figura 7.11. Já com os dados no período de baixa, no mesmo cenário, encontramos com o teste KS que existe normalidade nos dados para os dois índices IFIX e IBOV, conforme Figura 7.12. Desta forma, para os índices em altas usamos testes não paramétricos para comparação da média e testes paramétricos nos caso dos índices em baixa. Além disso, os resultados das Tabelas 7.5 e 7.6 nos mostram que em ambas as situações, de alta e de baixa, o índice IFIX obteve um retorno maior que o índice IBOV e risco menor, com uma única diferença em relação ao artigo [69], em que no período de alta dos índices o índice IFIX teve um retorno menor.

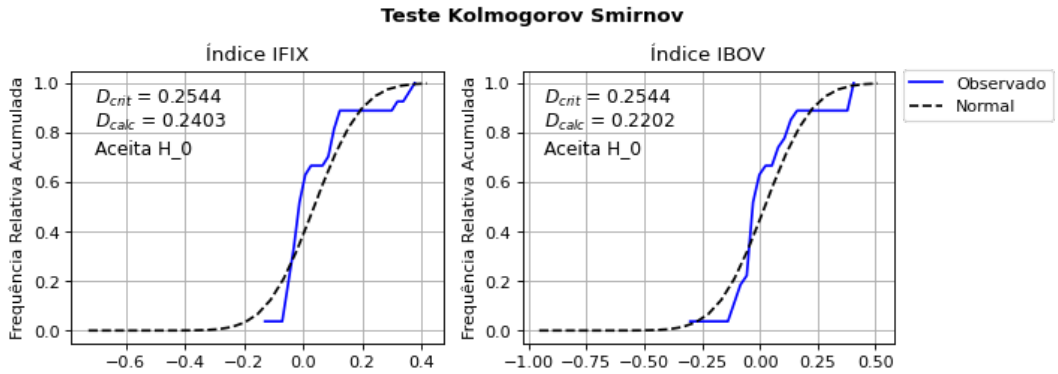

Figura 7.12: Gráficos dos resultados do teste de hipótese Kolmogorov-Smirnov para averiguar a suposição de normalidade considerando o período com IFIX e IBOV em baixa.
Fonte: O autor (2021).

Na análise dos resultados dos testes de comparação da média, com os testes adequados para cada um dos dois casos, obtemos pelo teste bicaudal Wilcoxon que os índices IFIX e IBOV possuem retornos significantemente iguais, isto é, $-1.960 = -z_{\alpha/2} < W_{calc} =$

$z_{calc} < z_{\alpha/2} = 1.960$, para quando os índices estão em alta, ver Figura 7.13. Já no outro caso, quando os índices estão em baixa, também obtemos que as médias são estatisticamente iguais, ou seja, $-2.052 = -t_{\alpha/2} < t_{calc} < t_{\alpha/2} = 2.052$, de acordo com a Figura 7.14, com nível de significância $\alpha = 0.05$ em ambos os testes. Logo, os retornos com ambos os índices IFIX e IBOV em alta são semelhantes, não havendo diferença significante em investir em um ou outro índice. Porém, isso não pode ser levado para a vida real de investimentos, pois nem sempre temos os dois índices com período de alta.

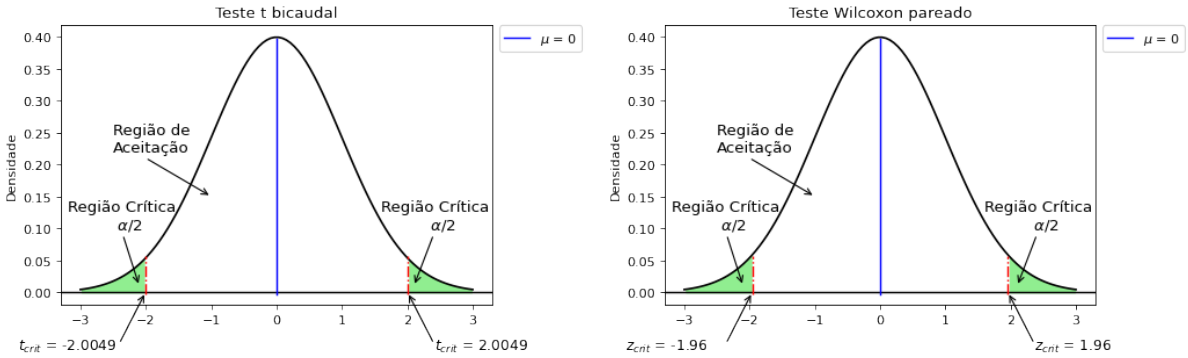

Figura 7.13: Gráficos dos testes de hipóteses t e Wilcoxon para comparação da média considerando período com IFIX e IBOV em alta.
Fonte: O autor (2021).

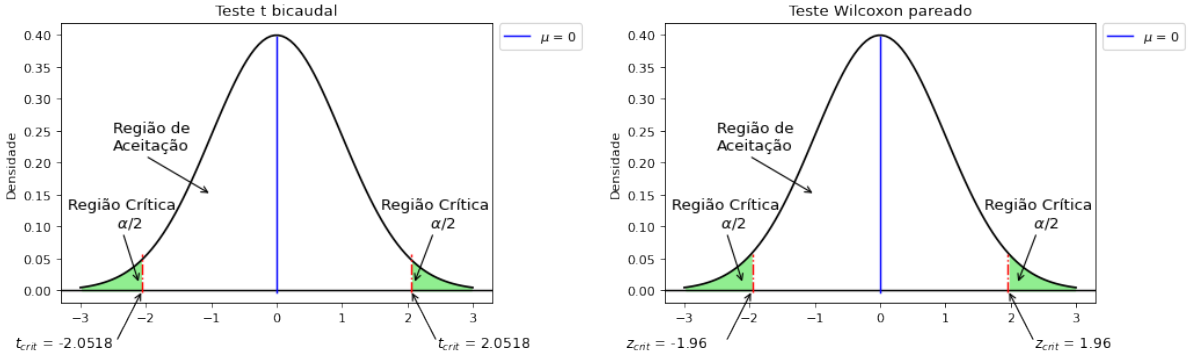

Figura 7.14: Gráficos dos testes de hipóteses t e Wilcoxon para comparação da média considerando período com IFIX e IBOV em baixa.
Fonte: O autor (2021).

Em seguida verificamos a homogeneidade, através dos resultados das Tabelas 7.5 e 7.6. Aferimos que a variância entre os índices no caso em que ambos estão em alta usando teste Levene não apresentaram ser significantemente iguais, em conformidade com as Figuras 7.15 e 7.16. Contudo, no caso em que ambos os índices estão em baixa usando teste F houve a aceitação da hipótese nula de homogeneidade de variâncias. Isso significa que nos meses em que ambos os índices estão em baixa, a dispersão se sucede de forma igual ou semelhante, ou seja, no caso de baixa dos dois índices a volatilidade se apresenta semelhante não tendo o IFIX uma volatilidade menor que o índice IBOV.

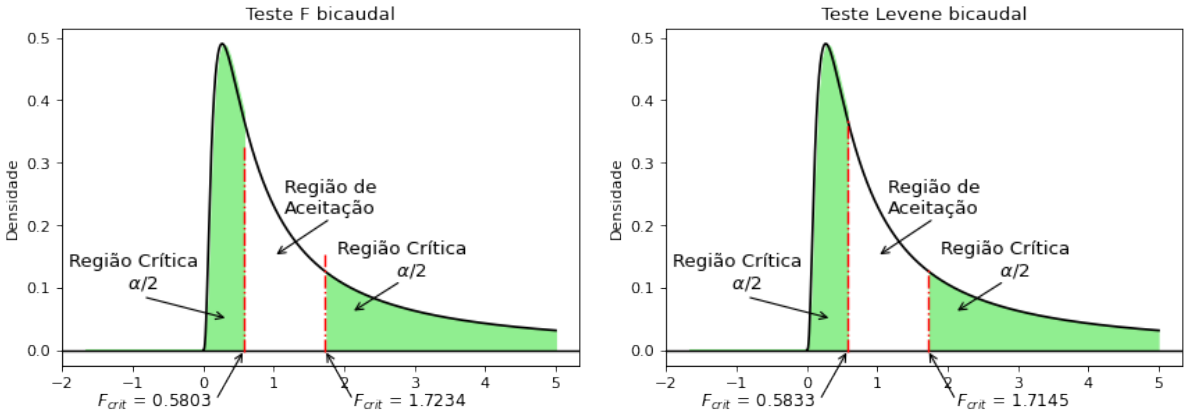

Figura 7.15: Gráficos dos testes de hipóteses F e Levene para comparação da variância considerando período com IFIX e IBOV em alta.
Fonte: O autor (2021).

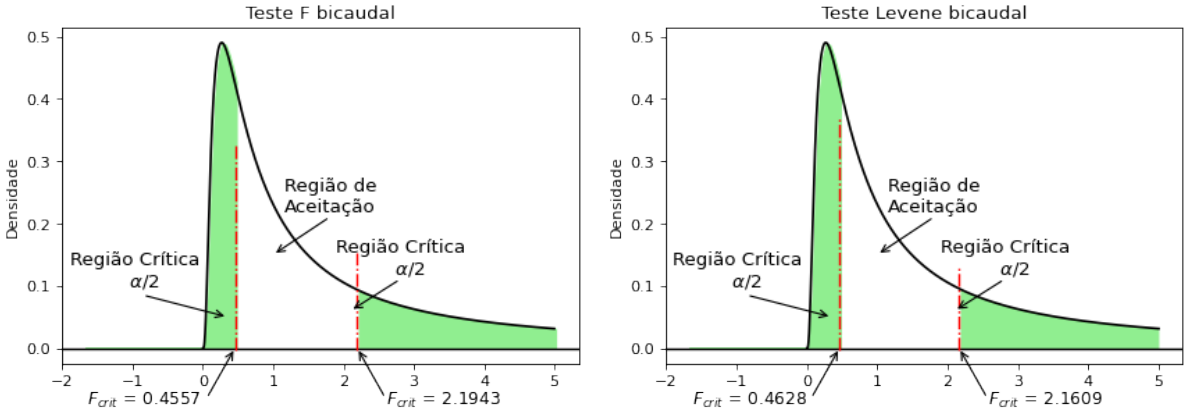

Figura 7.16: Gráficos dos testes de hipóteses F e Levene para comparação da variância considerando período com IFIX e IBOV em baixa.
Fonte: O autor (2021).

Mesmo os dados tendo aparente correlação, utilizamos os métodos para comparação da variância que requer o pressuposto de independência dos dados tanto dos dados log-retorno para usar o teste F, porque acreditamos que teríamos bons resultados. Contudo, para análises futuras aconselhamos a aplicação de algum teste de hipótese para verificar esse pressuposto de independência, além do pressuposto de normalidade que testamos para a utilização dos testes de hipóteses.

Tabela 7.5: Média, desvio-padrão, teste de normalidade, teste de igualdade de média e teste de igualdade de variância para os meses de alta para ambos os índices IBOV e IFIX.

Estatística ou Teste	**IFIX**	**IBOV**
Média	1.939%	0.992%
Desvio-Padrão	4.876%	8.702%
Normalidade ($p - valor$)	0.003	0.005

Continua na próxima página

Estatística ou Teste	IFIX	IBOV
Normalidade (D_{calc})	0.241	0.234
Igualdade de Média ($p - valor$)	-	0.797
Igualdade de Média (t_{calc}/W_{calc})	-	-0.482
Igualdade de Variância ($p - valor$)	-	0.015
Igualdade de Variância (F_{calc}/L_{calc})	-	6.085

Fonte: O autor (2021).

Tabela 7.6: Média, desvio-padrão, teste de normalidade, teste de igualdade de média e teste de igualdade de variância para os meses de baixa para ambos os índices IBOV e IFIX.

Estatística ou Teste	IFIX	IBOV
Média	3.704%	2.149%
Desvio-Padrão	12.470%	15.958%
Normalidade ($p - valor$)	0.074	0.125
Normalidade (D_{calc})	0.240	0.220
Igualdade de Média ($p - valor$)	-	0.514
Igualdade de Média (t_{calc}/W_{calc})	-	0.662
Igualdade de Variância ($p - valor$)	-	0.215
Igualdade de Variância (F_{calc}/L_{calc})	-	0.611

Fonte: O autor (2021).

Como desfecho da análise de comparação do índice IFIX com outros índices financeiros, em resumo concluímos que

- considerando todo o período, todos os índices comparados têm retornos estatisticamente iguais, mas com volatilidades diferentes,
- apenas utilizando período com IBOV em alta, o índice IDIV não tem retorno estatisticamente igual e a volatilidade se dá de maneira diferente do IFIX em comparação com os outros índices,
- no período com IBOV em baixa, somente o índice IMOB não possui retorno estatisticamente igual ao do índice IFIX e as variâncias continuam sendo estatisticamente diferentes;
- quando avaliamos no período com os índices IFIX e IBOV em alta, os resultados entre os índices IFIX e IBOV continuam os mesmos, com média estatisticamente igual e volatilidade diferente,
- no último caso, com período com os índices IFIX e IBOV em baixa, os retornos são estatisticamente iguais entre os índices IFIX e IBOV e averiguamos que as variâncias são iguais, o que mostra que a volatilidade ocorre de modo semelhante. Desta forma, os índices IBOV e IFIX apresentam frequência da oscilação dos valores de fechamento mensais de forma semelhante.

Considerações finais

Neste trabalho, realizamos um estudo para compreensão dos conceitos principais de fundos de investimentos imobiliários e análises para averiguar a performance destes através de dados coletados do Índice de Fundos Imobiliários (IFIX). Na parte teórica, exibimos e compreendemos os tipos de fundos imobiliários e a divisão por objetivo dos mesmos. Na parte numérica, conseguimos efetuar previsões de tendência para o futuro dos fundos através do índice IFIX, também da tendência do Ibovespa, e analisar o desempenho dos fundos, em relação ao retorno e risco, comparando o índice IFIX com outros índices que possuem relação com o meio imobiliário e ações. Para essa finalidade, aprendemos a utilizar na linguagem de programação Python testes de hipóteses para as comparações dos índices e métodos de Aprendizagem de Máquinas para realizar as previsões de h meses à frente.

A partir da análise de tendência, concluímos que a previsão de tendência de índices ou ações, que tratamos como um problema de classificação, é uma aplicação trabalhosa de replicarmos de forma prática, porque a análise de tendência usando o índice Ibovespa ou IFIX mostrou que os resultados dependem da quantidade de meses que desejamos prever à frente e, principalmente, do período dos dados utilizados, pois podemos ter influência de períodos de crise e de grande evolução. Dos resultados mostrados no Capítulo 6 percebemos, em geral, que os modelos de Aprendizagem de Máquinas apresentaram melhores medidas de avaliação nos casos com $h = 1$ e $h = 12$ dependendo do índice utilizado na análise, IFIX ou IBOV. Sendo que quando utilizamos apenas dados do índice, sempre o melhor modelo encontrado é utilizando $h = 1$, contudo o melhor modelo, em ambas as análises, foi com previsão de um ano à frente $h = 12$ que utiliza a retirada de variáveis com os resultados da ANOVA. Quando aplicamos apenas dados do índice, o modelo que deteve o melhor resultado foi a Regressão Logística, mas quando utilizamos dados de mercado e / ou *feature engineering* nenhum dos modelos presentou destaque, pois como vimos nos resultados, Regressão Logística, *XGBoost*, Máquina de Vetores Suporte e Rede Neural tiveram resultados como o melhor modelo em alguma das análises, em que fomos adicionando mais variáveis explicativas e/ou utilizando ANOVA para selecionais as variáveis mais significativas.

Comparando com a dissertação de mestrado [21], que seguimos para a análise de tendência, o que acrescentamos foi a utilização de dados de mercado e criação de novas variáveis através de *feature engineering*. Além disso, utilizamos apenas três meses de valores mais próximos do valor que desejamos prever, isso se mostrou um diferencial, pois sucedeu em melhores resultados em relação à acurácia, que na dissertação denominam por taxa de acerto. Contudo, como estamos com dados desbalanceados, não podemos utilizar essa medida para realizar a avaliação do modelo. Desta forma, empregamos outra métrica de avaliação, a acurácia balanceada. Caso pudéssemos comparar em relação à

acurácia, o melhor modelo para a análise com o índice IFIX seria o usando o modelo RN com $h = 12$ com a utilização de apenas os dados do índice, ou as outras situações, porque os resultados se mantiveram similares, com a acurácia de 82.61%. Já em relação ao índice IBOV, os melhores modelos foram RN com $h = 12$ nas situações com dados de mercado e ANOVA, e todos os dados mais ANOVA, mas quando estamos com dados do índice, de mercado e variáveis obtidas com *feature engineering* o melhor modelo foi o SVM, sendo que estes modelos apresentaram acurácia de 82.98%.

Comparando o índice IFIX com os índices IBOV, IDIV, SMLL e IMOB, no que diz respeito ao retorno e risco, concluímos que considerando todo o período sem separar em períodos com IBOV em alta e baixa (ABAB) e IBOV e IFIX em alta e baixa (AIAB), que os retornos ou média de ganho podem ser considerados iguais, o que significa que quando realizamos uma aplicação em qualquer um destes índices teremos retorno monetário similar, ou seja, não há diferença entre investir no índice IFIX e nos demais. Já em relação à volatilidade concluímos que os índices possuem volatilidades diferentes, o que significa que os riscos obtidos nas aplicações de cada um destes índices não podem ser considerados similares. Na análise com ABAB, a diferença com o resultado anterior pode ser vista somente no fato do índice IDIV e IMOB não apresentarem retornos similares em período de alta e baixa, respectivamente. Na última análise realizada, AIAB, com comparação entre IBOV e IFIX averiguamos que os retornos continuam similares, mas com a volatilidade acontecendo de forma igual ou similar. Este resultado pode ter relação com a crise causada pela COVID-19, no qual houve uma grande queda nos índices financeiros no período de Janeiro de 2020 a Abril de 2020 e posterior recuperação vagarosa que averiguamos pela Figura 1.1, pois uma parte dos dados que colocamos a mais possuem referência a esse período de crise.

Projetos futuros que podem melhorar as análises realizadas são os seguintes

1. Replicar a análise de tendência considerando os casos:

   (a) Duas novas análises: uma apenas utilizando dados dos meses em que houve crise na economia, com queda do índice, e outra com os meses sem, com a finalidade de sabermos se as análises continuam as mesmas e os resultados ficam mais estáveis. Contudo, como teremos menos dados, uma possibilidade é usar dados diários dos índices.

   (b) Para os fundos imobiliários e ações, com o intuito de verificarmos se os resultados se mantêm instáveis ou melhoram dependendo do fundo imobiliário, da ação, ou ainda do setor em que estão presentes.

   (c) Aplicando análises que utilizam fundos imobiliários ou ações, separadas por tipo ou setor, em vez dos índices Ibovespa e IFIX. Também aqui, fazer análises com e sem período de crise econômica.

   (d) Com e sem as opções anteriores, aplicando alguma técnica para balanceamento dos dados como, por exemplo, *Synthetic Minority Oversampling Technique* (SMOTE) e *Adaptive Synthetic Sampling* (ADASYN).

2. A análise de comparação entre os índices poderíamos refazer em duas análises. Uma com meses em que ocorreu crise econômica e outra em que não ocorreu.

Referências

[1] ANALYTICS VIDHYA. **Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python)**. Disponível em: https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/. Acesso em: 24 mar. 2021.

[2] AZANK, F.; GURGEL, G. K.. MEDIUM. **Dados Desbalanceados — O que são e como lidar com eles**. Disponível em: https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b. Acesso em: 26 out. 2021.

[3] B3. Disponível em: http://www.b3.com.br/pt_br/. Acesso em: 27 abr. 2020.

[4] BEVANS, R. **The p-value explained**. Disponível em: https://www.scribbr.com/statistics/p-value/. Acesso em: 18 set. 2021.

[5] BONESSO, D. **Estimação dos parâmetros do kernel em um classificador svm na classificação de imagens hiperespectrais em uma abordagem multiclasse**. 2013. Dissertação (Mestrado em Informática), Programa de Pós-Graduação em Sensoriamento Remoto, Porto Alegre. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/86168/000909969.pdf?sequence=1. Acesso em: 22 out. 2021.

[6] BROWN, M. B.; FORSYTHE, A. B. **Robust Tests for the Equality of Variances**. Journal of the American Statistical Association, 1974, pp. 364-367 Stanford University Press, 1960, pp. 278-292. Acesso em: 02 mai. 2021.

[7] CAMPAGNARO, R. FIIS. **8 motivos para investir em Fundos Imobiliários**. Disponível em: https://fiis.com.br/artigos/8-motivos-para-investir-em-fundos-imobiliarios/. Acesso em: 15 mar. 2020.

[8] CAMPAGNARO, R. FIIS. **10 indicadores importantes para investir em Fundos Imobiliários**. Disponível em: https://fiis.com.br/artigos/indicadores-fundos-imobiliarios/. Acesso em: 14 abr. 2020.

[9] CAMPAGNARO, R. FIIS. **ABL, área privativa, área BOMA...o que seriam tais conceitos?**. Disponível em: https://fiis.com.br/artigos/abl-area-privativa-area-boma-o-que-seriam-tais-conceitos/. Acesso em: 16 abr. 2020.

[10] CAMPAGNARO, R. FIIS. **Tipos de Fundos Imobiliários – Confira todas as opções para investir**. Disponível em: https://fiis.com.br/artigos/tipos-de-fundos-imobiliarios/. Acesso em: 03 fev. 2020.

[11] CARVALHO, P. L. d.; SOUSA, E.; CALLADO, A. L. C. **Indicadores de Desempenho da BM & FBovespa: Um Análise do Desempenho Financeiro dos Índices de Sustentabilidade frente aos demais Índice da bolsa.** Encontro Internacional sobre Gestão Empresarial e Meio Ambiente-ENGEMA. ISSN. pp. 2359–1048. 2016. Disponível em: http://engemausp.submissao.com.br/18/anais/arquivos/488.pdf Acesso em: 06 jul. 2021.

[12] CHEN, T.; GUESTRIN, C. **Xgboost: A scalable tree boosting system.** 2016. roceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794. Disponível em: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785. Acesso em: 21 out. 2021.

[13] CORRÊA, D. M. ATELIWARE. **Feature Engineering: Preparando dados para o aprendizado de máquina**. Disponível em: https://ateliware.com/blog/feature-engineering. Acesso em: 09 nov. 2021.

[14] CREDIT SUISSE HEDGING-GRIFFO. **CSHG Real Estate FII Julho 2021**. Disponível em: https://imobiliario.cshg.com.br/central-de-downloads/relatorios-periodicos/hgre/. Acesso em: 05 ago. 2021.

[15] DIEZ, D. M.; BARR, C. D.; CETINKAYA-RUNDERL, M. **OpenIntro statistics.** Edição 2. OpenIntro, 2012.

[16] DUBARD, C. MAGNETIS. **Entenda mais sobre as taxas de administração e de performance dos fundos**. Disponível em: https://blog.magnetis.com.br/taxas-de-administracao-e-de-performance/. Acesso em: 17 ago. 2021.

[17] DUBARD, C. MAGNETIS. **Índice Bovespa: o que é? Como esse indicador é calculado? Descubra aqui!**. Disponível em: https://blog.magnetis.com.br/o-que-e-indice-bovespa/. Acesso em: 21 out. 2020.

[18] ELEVEN FINANCIAL. **Como escolher um Fundo Imobiliário (FII)?**. Disponível em: https://elevenfinancial.com/como-escolher-fundo-imobiliario. Acesso em: 15 mar. 2021.

[19] FAYH, M. THE CAPITAL ADIVISOR. **O que é IFIX e Como Funciona o Índice de Fundos Imobiliários**. Disponível em: https://comoinvestir.thecap.com.br/o-que-e-ifix-como-funciona-indice-fundos-imobiliarios/. Acesso em: 21 out. 2020.

[20] FERREIRA, C. D. S. **Cap. 5. Testes de Hipóteses**. Disponível em: https://www.ufjf.br/clecio_ferreira/files/2012/04/Cap5-Testes-de-hipoteses-Parte-13.pdf. Acesso em: 16 set. 2021.

[21] FINKLER, A. C. **Aprendizem de Máquina Aplicada à Previsão dos Movimentos do Ibovespa.** 2017. Dissertação (Mestrado em Matemática), Programa de Pós-Graduação em Matemática, Curitiba. Disponível em: https://acervodigital.ufpr.br/bitstream/handle/1884/49395/R%20-%20D%20-%20ALINE%20CRISTIANE%20FINKLER.pdf?sequence=1&isAllowed=y. Acesso em: 02 jul. 2021.

[22] FRATA, D. S. X.; BARCELLOS, D. K. CAPITALIZO. **Índice Small Cap (SMLL): o que é e como investir?**. Disponível em: https://capitalizo.com.br/indice-small-cap-smll-o-que-e-e-como-investir/. Acesso em: 21 out. 2020.

[23] FURLETTI, D. Í. R.; VASCONCELOS. I. M. P.; FREITAS, R. **Número-Índice: uma visão geral**. Disponível em: https://www.sinduscon-mg.org.br/site/arquivos/up/economica/Numero_Indice.pdf. Sindicato da Indústria da Construção Civil no Estado de Minas Gerais - Sinduscon-MG. Acesso em: 15 out. 2021.

[24] GENIAL INVESTIMENTOS. **O que é CRI? Aprenda tudo sobre o título isento de IR**. Disponível em: https://blog.genialinvestimentos.com.br/o-que-e-cri/. Acesso em: 29 abr. 2020.

[25] GENIAL INVESTIMENTOS. **Os 5 principais tipos de fundos imobiliários quanto às suas estratégias**. Disponível em: https://blog.genialinvestimentos.com.br/tipos-de-fundos-imobiliarios/. Acesso em: 31 abr. 2020.

[26] GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.** Edição 2. O'Reilly Media, 2019.

[27] GUIMARÃES, P. R. B. **Estatística não-paramétrica**. Disponível em: https://docs.ufpr.br/~prbg/public_html/ce050/aluno%202015%20np.pdf. Curitiba. Apostila (Disciplina Estatística Não Paramétrica)–Curso de Estatística, Setor de Ciências Exatas, Universidade Federal do Paraná. Acesso em: 16 jan. 2021.

[28] HECKERT, N. A.; JAMES, J. F. **NIST/SEMATECH e-Handbook of Statistical Methods; Chapter 1: Exploratory Data Analysis**. Disponível em: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm. Acesso em: 03 mai. 2021.

[29] HOMEM, W. L. **Apostila de Machine Learning**. Disponível em: https://petmecanica.ufes.br/sites/petengenhariamecanica.ufes.br/files/field/anexo/apostila_do_minicurso_de_machine_learning.pdf. Acesso em: 25 out. 2021.

[30] IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Índice Nacional de Preços ao Consumidor Amplo**. Disponível em: https://sidra.ibge.gov.br/tabela/1737. Acesso em: 25 jul. 2021.

[31] KOWALCZYK, A. **SVM - Understanding the math - Part 1 - The margin**. Disponível em: https://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/. Acesso em: 21 out. 2021.

[32] INFOMONEY. **LCI e LCA: guia completo para começar a investir**. Disponível em: https://www.infomoney.com.br/guias/lci-lca/. Acesso em: 29 abr. 2020.

[33] LEVENE, H. **In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds**. Stanford University Press, 1960, pp. 278-292. Acesso em: 02 mai. 2021.

[34] MACHADO, G. V. MEDIUM. **Guia de Redes Neurais Artificiais — Parte 2**. Disponível em: https://medium.com/@guilhermevallimmachado/guia-de-redes-neurais-artificiais-parte-2-28bbdfbee1dd. Acesso em: 25 out. 2021.

[35] MCCULLOCH, W. S.; PITTS, W. H. **A logical calculus of the ideas immanent in nervous activity.**. Bulletin of Mathematical Biophysics, 1943, n. 7, p. 115 – 133. Disponível em https://link.springer.com/article/10.1007%2FBF02478259. Acesso em: 25 out. 2021.

[36] MEDEIROS, R. EU QUERO INVESTIR. **Entenda como funciona o Índice Small Cap e porque você deveria ficar de olho nestas ações**. Disponível em: https://www.euqueroinvestir.com/indice-small-cap/. Acesso em: 21 out. 2020.

[37] MELO, C. SIGMOIDAL.AI. **Redes Neurais Multicamadas com Python e Keras**. Disponível em: https://sigmoidal.ai/redes-neurais-python-keras-2/. Acesso em: 25 out. 2021.

[38] MELONI, R. **Classificação de Imagens de Sensoriamento Remoto usando SVM**. 2017. Dissertação (Mestrado em Informática), Programa de Pós-Graduação em Informática, Rio de Janeiro. Disponível em: https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=31439@1. Acesso em: 22 out. 2021.

[39] MILLER, M. TOWARDS DATA SCIENCE. **The Basics: Logistic Regression and Regularization**. Disponível em: https://towardsdatascience.com/the-basics-logistic-regression-and-regularization-828b0d2d206c. Acesso em: 19 out. 2021.

[40] MIYAKI, K. MEDIUM2. **Time Series Split with Scikit-learn**. Disponível em: https://medium.com/keita-starts-data-science/time-series-split-with-scikit-learn-74f5be38489e. Acesso em: 18 out. 2021.

[41] MORDE, V. TOWARDS DATA SCIENCE. **XGBoost Algorithm: Long May She Reign!**. Disponível em: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. Acesso em: 21 out. 2021.

[42] NOGARE, D. **Performance de Machine Learning – Matriz de Confusão**. Disponível em: https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/. Acesso em: 04 out. 2021.

[43] NORDSTOKKE, D. W.; ZUMBO, B. D. **A Cautionary Tale about Levene's Tests for Equal Variances.**. Journal of Educational Research & Policy Studies, 2007, pp. 1-14. Disponível em https://files.eric.ed.gov/fulltext/EJ809430.pdf. Acesso em: 03 mai. 2021.

[44] PERPETUAL ENIGMA. **Kernel Functions For Machine Learning** . Disponível em: https://prateekvjoshi.com/2012/09/01/kernel-functions-for-machine-learning/. Acesso em: 22 out. 2020.

[45] PORTAL ACTION. **3 - TESTE DE WILCOXON PAREADO**. Disponível em: https://web.archive.org/web/20210226001931/http://www.portalaction.com.br/tecnicas-nao-parametricas/teste-de-wilcoxon-pareado. Acesso em: 15 jan. 2021.

[46] PORTAL ACTION. **5.1 - Retornos**. Disponível em: https://web.archive.org/web/20210410151821/http://www.portalaction.com.br/series-temporais/51-retornos. Acesso em: 10 mai. 2020.

[47] PORTAL ACTION. **5.6 - TESTE PARA COMPARAÇÃO DE DUAS VARIÂNCIAS (TESTE F)**. Disponível em: https://web.archive.org/web/20210121171114/http://www.portalaction.com.br/inferencia/56-teste-para-comparacao-de-duas-variancias-teste-f. Acesso em: 02 out. 2020.

[48] PORTAL ACTION. **5.8 - TESTE T PAREADO**. Disponível em: https://web.archive.org/web/20210122212546/http://www.portalaction.com.br/inferencia/58-teste-t-pareado. Acesso em: 02 out. 2020.

[49] REBELO, L. D. T. **Avaliação automática do resultado estético do tratamento conservador do cancro da mama**. 2008. Dissertação de Mestrado em Integrado em Engenharia Electrotécnica e de Computadores Major Telecomunicações, Porto, Portugal. Disponível em: https://repositorio-aberto.up.pt/bitstream/10216/60215/2/Texto%20integral.pdf. Acesso em: 16 dez. 2021.

[50] REIS, T. SUNO RESEARCH. **Fundo monoativo: como funciona? Vale a pena investir?**. Disponível em: https://www.suno.com.br/artigos/fundo-monoativo/. Acesso em: 03 out. 2021.

[51] REIS, T. SUNO RESEARCH. **IDIV: entenda como funciona o Índice de Dividendos da B3**. Disponível em: https://www.sunoresearch.com.br/artigos/idiv/. Acesso em: 20 out. 2020.

[52] REIS, T. SUNO RESEARCH. **Imob: entenda como funciona o Índice Imobiliário da B3**. Disponível em: https://www.sunoresearch.com.br/artigos/imob/. Acesso em: 22 out. 2020.

[53] REIS, T. SUNO RESEARCH. **Valor patrimonial: saiba como calcular e analisar esse indicador**. Disponível em: https://www.suno.com.br/artigos/valor-patrimonial/. Acesso em: 29 abr. 2020.

[54] RICO INVESTIMENTOS. **Qual o Seu Perfil de Investidor - Conservador, Moderado ou Arrojado?**. Disponível em: https://blog.rico.com.vc/perfil-de-investidor. Acesso em: 03 fev. 2020.

[55] RICO INVESTIMENTOS. **Renda Fixa x Renda Variável: Diferenças e Como Escolher a Melhor**. Disponível em: https://blog.rico.com.vc/renda-fixa-renda-variavel. Acesso em: 03 fev. 2020.

[56] ROSENBLATT, F. **The perceptron: a probabilistic model for information storage and organization in the brain.**. Psychological review, 1958, pp. 386.

Disponível em https://psycnet.apa.org/record/1959-09865-001. Acesso em: 25 out. 2021.

[57] SEMOLINI, R. **Support vector machines, inferência transdutiva e o problema de classificação.** 2002. Dissertação (Mestrado em Engenharia Elétrica)– Programa de Pós-graduação da Faculdade de Engenharia Elétrica e de Computação da Universidade de Campinas. Disponível em: https://www.dca.fee.unicamp.br/~vonzuben/theses/semolini_mest/semolini_tese_mest.pdf. Acesso em: 22 out. 2021.

[58] ROSSUM, G. V.; DRAKE, F. L. **Python 3 Reference Manual.** Haarlem: CreateSpace, 2009.

[59] SANTANA, F. MINERANDO DADOS. **Árvores de Decisão (Projeto passo a passo)**. Disponível em: https://minerandodados.com.br/arvores-de-decisao-conceitos-e-aplicacoes/. Acesso em: 20 out. 2021.

[60] SANTOS, G. C. **Algoritmos de machine learning para previsão de ações da B3.** 2020. Dissertação (Mestrado em Matemática), Programa de Pós-Graduação da Faculdade de Engenharia Elétrica de Uberlândia. Disponível em: http://clyde.dr.ufu.br/bitstream/123456789/29897/7/AlgoritmosMachineLearning.pdf. Acesso em: 06 jul. 2021.

[61] SANTOS, G. MEDIUM. **Estatística para Seleção de Atributos**. Disponível em: https://medium.com/data-hackers/estat%C3%ADstica-para-sele%C3%A7%C3%A3o-de-atributos-81bdc274dd2c. Acesso em: 15 out. 2021.

[62] SCIKIT LEARN. **1.4. Support Vector Machines**. Disponível em: https://scikit-learn.org/stable/modules/svm.html. Acesso em: 21 out. 2021.

[63] SCIKIT LEARN. **3.1. Cross-validation: evaluating estimator performance**. Disponível em: https://scikit-learn.org/stable/modules/cross_validation.html. Acesso em: 18 out. 2021.

[64] SCIKIT LEARN. **sklearn.feature_selection.SelectPercentile**. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html. Acesso em: 20 dez. 2021.

[65] SCIKIT LEARN. **sklearn.feature_selection.f_classif**. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html. Acesso em: 20 dez. 2021.

[66] SCIKIT LEARN. **sklearn.metrics.confusion_matrix**. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. Acesso em: 18 out. 2021.

[67] SCIKIT LEARN. **sklearn.preprocessing.MinMaxScaler**. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Acesso em: 25 jul. 2021.

[68] SCIKIT LEARN. **Visualizing cross-validation behavior in scikit-learn**. Disponível em: https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html#sphx-glr-auto-examples-model-selection-plot-cv-indices-py. Acesso em: 18 out. 2021.

[69] SERRA, R. G.; MORAES, A. V. d. **Comparação do Risco-Retorno do IFIX com IBOVESPA, IDIV, SMLL E IMOB**. Disponível em:https://www.researchgate.net/profile/Ricardo-Serra/publication/321978287_Comparacao_do_risco-retorno_do_IFIX_com_IBOVESPA_IDIV_SMLL_e_IMOB/links/5a74a4700f7e9b20d4923ef1/Comparacao-do-risco-retorno-do-IFIX-com-IBOVESPA-IDIV-SMLL-e-IMOB.pdf. 2017. Acesso em: 19 abr. 2020.

[70] SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétrica para ciências do comportamento.** Edição 2. Porto Alegre: Artmed Editora, 2006.

[71] SILVA, A. d. N. **Utilização de algoritmo com rede neural artificial na validação de padrões de comportamento do C. elegans**. 2017. Centro Universitário SENAI CIMATEC. Disponível em: http://repositoriosenaiba.fieb.org.br/bitstream/fieb/900/1/Alexandre%20do%20Nascimento%20Silva.pdf. Acesso em: 25 out. 2021.

[72] SMITH, TIM. INVESTOPEDIA. **Qualitative Analysis**. Disponível em: https://www.investopedia.com/terms/q/qualitativeanalysis.asp. Acesso em: 17 ago. 2021.

[73] SONDEREGGER, D. L.; BUSCAGLIA, R. **Introduction to Statistical Methodology**. Second Edition. Disponível em: https://bookdown.org/dereksonderegger/570/. Acesso em: 15 out. 2021.

[74] SPERANDIO, B. FIIS. **Como calcular taxa de vacância – É muito simples.**. Disponível em: https://fiis.com.br/artigos/como-calcular-taxa-de-vacancia/. Acesso em: 29 abr. 2020.

[75] SPERANDIO, B. FIIS. **O que é Cap Rate - Saiba o que é antes de comprar um imóvel**. Disponível em: https://fiis.com.br/artigos/o-que-e-cap-rate/. Acesso em: 01 mai. 2020.

[76] SUNO RESEARCH. **Aprenda a Investir em Fundos Imobiliários**. Disponível em: https://lp.suno.com.br/ebook-investindo-em-fiis/. Acesso em: 07 mar. 2020.

[77] TUTORIAL AND EXAMPLE. **Random Forest Algorithm for Machine Learning**. Disponível em: https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/. Acesso em: 24 mar. 2021.

[78] TYAGI, N. **Understanding the Gini Index and Information Gain in Decision Trees**. Disponível em: https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8. Acesso em: 20 out. 2021.

[79] YAHOO! FINANCE. Disponível em: `https://finance.yahoo.com/`. Acesso em: 15 abr. 2020.

Anexo A - Tabelas dos resultados da análise de tendência usando métodos de Machine Learning - índice Ibovespa

Neste Anexo apresentamos as tabelas dos resultados da análise de tendência usando ML na aplicação com os dados do índice Ibovespa, em relação aos métodos ML e janela de previsão h meses à frente. A Tabela A.1 compreende a análise somente com o índice IBOV, a Tabela A.2 contém os resultados utilizando dados do índice e dados de mercado considerados utilizando 90% das variáveis mais significativas, na Tabela A.3 temos os resultados relativos ao incremento de variáveis usando *feature engineering* sem aplicação da ANOVA e na Tabela A.4 exibimos os resultados utilizando 90% das variáveis mais significativas (índice, dados de mercado e *feature engineering*) obtidas pela ANOVA. Os três melhores resultados em relação à medida de avaliação acurácia balanceada estão em negrito nas tabelas.

Tabela A.1: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IBOV sem dados de mercado sem *feature engineering*.

Métodos	**h**	**Acurácia**	**Acurácia Balanceada**	**Precisão**	**Recall**	**F1-Score**
RL		67.35%	**68.37%**	0.80	0.65	0.71
RF		38.78%	51.61%	1.00	0.03	0.06
XGB	1	38.78%	51.61%	1.00	0.03	0.06
SVM		40.82%	40.41%	0.54	0.42	0.47
RN		40.82%	53.23%	1.00	0.06	0.12
RL		45.83%	**55.60%**	0.80	0.34	0.48
RF		29.17%	51.43%	1.00	0.03	0.06
XGB	3	29.17%	51.43%	1.00	0.03	0.06
SVM		50.00%	**53.63%**	0.76	0.46	0.57
RN		29.17%	51.43%	1.00	0.03	0.06
RL		27.08%	52.70%	1.00	0.05	0.10
RF		27.08%	52.70%	1.00	0.05	0.10
XGB	6	25.00%	51.35%	1.00	0.03	0.05
SVM		41.67%	43.00%	0.71	0.41	0.52
RN		22.92%	50.00%	0.00	0.00	0.00

Continua na próxima página

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	12	74.47%	44.87%	0.81	0.90	0.85
RF		17.02%	50.00%	0.00	0.00	0.00
XGB		17.02%	50.00%	0.00	0.00	0.00
SVM		31.91%	19.23%	0.65	0.38	0.48
RN		17.02%	50.00%	0.00	0.00	0.00

Fonte: O autor (2021).

Tabela A.2: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IBOV com dados de mercado sem *feature engineering*, usando 90% das variáveis obtidas com a ANOVA.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	36.73%	50.00%	0.00	0.00	0.00
RF		57.14%	**55.65%**	0.68	0.61	0.64
XGB		57.14%	54.48%	0.67	0.65	0.66
SVM		63.27%	50.00%	0.63	1.00	0.78
RN		63.27%	50.00%	0.63	1.00	0.78
RL	3	72.92%	50.00%	0.73	1.00	0.84
RF		72.92%	52.42%	0.74	0.97	0.84
XGB		47.92%	42.53%	0.68	0.54	0.60
SVM		52.08%	**64.73%**	0.93	0.37	0.53
RN		72.92%	50.00%	0.73	1.00	0.84
RL	6	77.08%	50.00%	0.77	1.00	0.87
RF		77.08%	50.00%	0.77	1.00	0.87
XGB		66.67%	43.24%	0.74	0.86	0.80
SVM		77.08%	50.00%	0.77	1.00	0.87
RN		22.92%	50.00%	0.00	0.00	0.00
RL	12	82.98%	50.00%	0.83	1.00	0.91
RF		19.15%	51.28%	1.00	0.03	0.05
XGB		48.94%	34.46%	0.76	0.56	0.65
SVM		65.96%	**79.49%**	1.00	0.59	0.74
RN		82.98%	50.00%	0.83	1.00	0.91

Fonte: O autor (2021).

Tabela A.3: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IBOV com dados de mercado e *feature engineering*.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	44.90%	55.29%	0.83	0.16	0.27
RF		48.98%	52.69%	0.67	0.39	0.49
XGB		42.86%	50.18%	0.64	0.23	0.33
SVM		36.73%	50.00%	0.00	0.00	0.00

Continua na próxima página

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RN	1	51.02%	55.47%	0.71	0.39	0.50
RL	3	54.17%	56.48%	0.78	0.51	0.62
RF		75.00%	**70.77%**	0.85	0.80	0.82
XGB		62.50%	47.69%	0.72	0.80	0.76
SVM		37.50%	54.73%	0.86	0.17	0.29
RN		72.92%	50.00%	0.73	1.00	0.84
RL	6	81.25%	**59.09%**	0.80	1.00	0.89
RF		79.17%	54.55%	0.79	1.00	0.88
XGB		77.08%	50.00%	0.77	1.00	0.87
SVM		77.08%	50.00%	0.77	1.00	0.87
RN		22.92%	50.00%	0.00	0.00	0.00
RL	12	82.98%	50.00%	0.83	1.00	0.91
RF		19.15%	51.28%	1.00	0.03	0.05
XGB		53.19%	**71.79%**	1.00	0.44	0.61
SVM		82.98%	50.00%	0.83	1.00	0.91
RN		17.02%	50.00%	0.00	0.00	0.00

Fonte: O autor (2021).

Tabela A.4: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IBOV com dados de mercado, *feature engineering*, usando 90% das variáveis obtidas com a ANOVA.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	44.90%	54.12%	0.75	0.19	0.31
RF		51.02%	54.30%	0.68	0.42	0.52
XGB		51.02%	48.48%	0.62	0.58	0.60
SVM		46.94%	52.24%	0.67	0.32	0.43
RN		36.73%	50.00%	0.00	0.00	0.00
RL	3	72.92%	50.00%	0.73	1.00	0.84
RF		47.92%	49.78%	0.73	0.46	0.56
XGB		60.42%	51.10%	0.74	0.71	0.72
SVM		41.67%	**57.58%**	0.89	0.23	0.36
RN		72.92%	50.00%	0.73	1.00	0.84
RL	6	79.17%	**54.55%**	0.79	1.00	0.88
RF		77.08%	50.00%	0.77	1.00	0.87
XGB		77.08%	50.00%	0.77	1.00	0.87
SVM		77.08%	50.00%	0.77	1.00	0.87
RN		22.92%	50.00%	0.00	0.00	0.00
RL	12	82.98%	50.00%	0.83	1.00	0.91
RF		17.02%	50.00%	0.00	0.00	0.00
XGB		57.45%	54.49%	0.85	0.59	0.70
SVM		63.83%	**78.21%**	1.00	0.56	0.72

Continua na próxima página

Métodos	**h**	**Acurácia**	**Acurácia Balanceada**	**Precisão**	**Recall**	**F1-Score**
RN	12	17.02%	50.00%	0.00	0.00	0.00

Fonte: O autor (2021).

Anexo B - Tabelas dos resultados da análise de tendência usando métodos de Machine Learning - índice IFIX

Neste Anexo exibimos tabelas resultantes da análise de tendência usando ML, referentes a aplicação com dados do índice IFIX. A Tabela B.1 tem referência à análise a utilização apenas dados do índice IFIX, a Tabela B.2 compreende a análise com a adição dos dados de mercado sem *feature engineering*, a Tabela B.3 têm a análise com *feature engineering* a mais e a Tabela B.4 os resultados são da análise com utilização de 95% das variáveis mais significativas (índice, dados de mercado e *feature engineering*) pela ANOVA. Os três melhores resultados em relação à medida de avaliação acurácia balanceada estão em negrito nas tabelas.

Tabela B.1: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IFIX sem dados de mercado sem *feature engineering*.

Métodos	**h**	**Acurácia**	**Acurácia Balanceada**	**Precisão**	**Recall**	**F1-Score**
RL	1	68.00%	**62.85%**	0.72	0.81	0.76
RF		64.00%	50.00%	0.64	1.00	0.78
XGB		64.00%	50.00%	0.64	1.00	0.78
SVM		64.00%	50.00%	0.64	1.00	0.78
RN		48.00%	49.65%	0.64	0.44	0.52
RL	3	75.00%	**54.74%**	0.81	0.89	0.85
RF		79.17%	50.00%	0.79	1.00	0.88
XGB		20.83%	50.00%	0.00	0.00	0.00
SVM		79.17%	50.00%	0.79	1.00	0.88
RN		70.83%	44.74%	0.77	0.89	0.83
RL	6	66.67%	44.44%	0.73	0.89	0.80
RF		25.00%	50.00%	0.00	0.00	0.00
XGB		25.00%	50.00%	0.00	0.00	0.00
SVM		75.00%	50.00%	0.75	1.00	0.86
RN		75.00%	50.00%	0.75	1.00	0.86
RL	12	39.13%	33.55%	0.73	0.42	0.53
RF		82.61%	50.00%	0.83	1.00	0.90
XGB		82.61%	50.00%	0.83	1.00	0.90
SVM		82.61%	50.00%	0.83	1.00	0.90

Continua na próxima página

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RN	12	82.61%	**50.00%**	0.83	1.00	0.90

Fonte: O autor (2021).

Tabela B.2: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IFIX com dados de mercado sem *feature engineering*.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	64.00%	**52.43%**	0.65	0.94	0.77
RF		60.00%	46.88%	0.63	0.94	0.75
XGB		56.00%	48.61%	0.63	0.75	0.69
SVM		64.00%	50.00%	0.64	1.00	0.78
RN		64.00%	**64.58%**	0.77	0.63	0.69
RL	3	70.83%	44.74%	0.77	0.89	0.83
RF		45.83%	51.05%	0.80	0.42	0.55
XGB		33.33%	**57.89%**	1.00	0.16	0.27
SVM		79.17%	50.00%	0.79	1.00	0.88
RN		79.17%	50.00%	0.79	1.00	0.88
RL	6	62.50%	41.67%	0.71	0.83	0.77
RF		20.83%	25.00%	0.43	0.17	0.24
XGB		20.83%	41.67%	0.00	0.00	0.00
SVM		75.00%	50.00%	0.75	1.00	0.86
RN		75.00%	50.00%	0.75	1.00	0.86
RL	12	82.61%	50.00%	0.83	1.00	0.90
RF		82.61%	50.00%	0.83	1.00	0.90
XGB		78.26%	47.37%	0.82	0.95	0.88
SVM		82.61%	50.00%	0.83	1.00	0.90
RN		82.61%	50.00%	0.83	1.00	0.90

Fonte: O autor (2021).

Tabela B.3: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IFIX com dados de mercado e *feature engineering*.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	56.00%	**55.90%**	0.69	0.56	0.62
RF		60.00%	49.31%	0.64	0.88	0.74
XGB		60.00%	49.31%	0.64	0.88	0.74
SVM		64.00%	50.00%	0.64	1.00	0.78
RN		64.00%	50.00%	0.64	1.00	0.78
RL	3	70.83%	44.74%	0.77	0.89	0.83
RF		79.17%	50.00%	0.79	1.00	0.88
XGB		29.17%	47.89%	0.75	0.16	0.26
SVM		79.17%	50.00%	0.79	1.00	0.88

Continua na próxima página

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RN	3	79.17%	50.00%	0.79	1.00	0.88
RL	6	66.67%	44.44%	0.73	0.89	0.80
RF		45.83%	47.22%	0.73	0.44	0.55
XGB		25.00%	44.44%	0.50	0.06	0.10
SVM		75.00%	50.00%	0.75	1.00	0.86
RN		37.50%	47.22%	0.71	0.28	0.40
RL	12	82.61%	**50.00%**	0.83	1.00	0.90
RF		82.61%	**50.00%**	0.83	1.00	0.90
XGB		82.61%	**50.00%**	0.83	1.00	0.90
SVM		82.61%	**50.00%**	0.83	1.00	0.90
RN		82.61%	**50.00%**	0.83	1.00	0.90

Fonte: O autor (2021).

Tabela B.4: Resultado da previsão para h meses usando métodos de Machine Learning - Índice IFIX com dados de mercado, *feature engineering* e 95% das variáveis obtidas com a ANOVA.

Métodos	h	Acurácia	Acurácia Balanceada	Precisão	Recall	F1-Score
RL	1	56.00%	**53.47%**	0.67	0.63	0.65
RF		60.00%	49.31%	0.64	0.88	0.74
XGB		64.00%	**54.86%**	0.67	0.88	0.76
SVM		64.00%	50.00%	0.64	1.00	0.78
RN		44.00%	46.53%	0.60	0.38	0.46
RL	3	70.83%	44.74%	0.77	0.89	0.83
RF		79.17%	50.00%	0.79	1.00	0.88
XGB		29.17%	47.89%	0.75	0.16	0.26
SVM		79.17%	50.00%	0.79	1.00	0.88
RN		79.17%	50.00%	0.79	1.00	0.88
RL	6	66.67%	44.44%	0.73	0.89	0.80
RF		45.83%	47.22%	0.73	0.44	0.55
XGB		16.67%	27.78%	0.25	0.06	0.09
SVM		75.00%	50.00%	0.75	1.00	0.86
RN		66.67%	44.44%	0.73	0.89	0.80
RL	12	82.61%	50.00%	0.83	1.00	0.90
RF		82.61%	50.00%	0.83	1.00	0.90
XGB		91.30%	**75.00%**	0.90	1.00	0.95
SVM		82.61%	50.00%	0.83	1.00	0.90
RN		82.61%	50.00%	0.83	1.00	0.90

Fonte: O autor (2021).

Anexo C - Tabelas das variáveis mais significativas obtidas através da ANOVA.

Neste Anexo apresentamos as tabelas da ANOVA com as variáveis mais significativas dos dados da análise de tendência com os índices IBOV e IFIX. Estas tabelas contém os resultados do modelo com todos os dados: índice, dados de mercado e *feature engineering*, conforme as análises para cada um dos índices IBOV e IFIX.

Tabela C.1: Variáveis mais significativas usando ANOVA com dados do índice, mercado e *feature engineering* - índice IBOV - parte 1

Variável	**h**	**F Score**	**P-Valor**	**Variável**	**h**	**F Score**	**P-Valor**
variável 1	1	7.1425	0.0080	variável 6	3	10.5937	0.0013
IPCA mês t-0		2.7121	0.1009	dolar mês t-1		9.0484	0.0029
variável 6		1.4544	0.2290	dolar mês t-0		8.3506	0.0042
dolar mês t-1		1.1596	0.2826	variável 3		8.2619	0.0044
mês t-1		1.0700	0.3020	dolar mês t-2		7.8406	0.0055
ouro mês t-1		0.9843	0.3221	variável 8		3.1243	0.0784
dolar mês t-0		0.9342	0.3348	ouro mês t-0		1.8493	0.1752
ouro mês t-0		0.8492	0.3577	ouro mês t-1		1.7123	0.1919
variável 4		0.7914	0.3746	variável 5		1.6705	0.1974
ouro mês t-2		0.7010	0.4033	variável 4		1.5933	0.2081
mês t-0		0.6944	0.4055	ouro mês t-2		1.5316	0.2171
variável 3		0.6354	0.4262	IPCA mês t-2		1.4595	0.2282
IPCA mês t-1		0.4521	0.5020	IPCA mês t-1		1.3202	0.2517
variável 2		0.4373	0.5091	mês t-1		0.9892	0.3210
dolar mês t-2		0.4325	0.5114	variável 2		0.9386	0.3336
variável 5		0.3950	0.5303	mês t-2		0.9076	0.3417
variável 8		0.3620	0.5479	variável 1		0.5515	0.4584
IPCA mês t-2		0.2813	0.5963	mês t-0		0.5424	0.4622
mês t-2		0.2253	0.6354	IPCA mês t-0		0.5311	0.4669
variável 7		0.0474	0.8278	variável 7		0.0002	0.9881

Fonte: O autor (2021).

Tabela C.2: Variáveis mais significativas usando ANOVA com dados do índice, mercado e *feature engineering* - índice IBOV - parte 2

Variável	**h**	**F Score**	**P-Valor**	**Variável**	**h**	**F Score**	**P-Valor**
dolar mês t-0	6	23.2966	0.0000	variável 6	12	38.0803	0.0000
dolar mês t-1		22.8547	0.0000	dolar mês t-0		32.7243	0.0000
variável 3		21.9513	0.0000	dolar mês t-1		30.0014	0.0000
variável 6		21.8820	0.0000	variável 3		29.1712	0.0000
dolar mês t-2		21.3690	0.0000	dolar mês t-2		28.5386	0.0000
ouro mês t-0		4.7072	0.0310	ouro mês t-0		13.2461	0.0003
ouro mês t-1		4.6491	0.0321	ouro mês t-1		13.0831	0.0004
variável 4		4.3629	0.0378	variável 4		12.9755	0.0004
ouro mês t-2		4.2127	0.0412	ouro mês t-2		12.887	0.0004
mês t-2		1.8941	0.1701	mês t-0		7.3524	0.0072
variável 2		1.8319	0.1772	mês t-1		6.8591	0.0094
mês t-1		1.6840	0.1957	variável 2		6.6653	0.0105
mês t-0		1.3553	0.2455	mês t-2		6.5192	0.0113
variável 5		0.9460	0.3318	variável 1		3.7844	0.0530
IPCA mês t-1		0.9071	0.3419	variável 8		2.228	0.1369
IPCA mês t-2		0.7594	0.3844	variável 5		1.3965	0.2385
IPCA mês t-0		0.7380	0.3912	IPCA mês t-2		1.3241	0.2511
variável 7		0.4959	0.4820	IPCA mês t-1		0.9496	0.3309
variável 1		0.1965	0.6580	IPCA mês t-0		0.7934	0.3740
variável 8		0.0065	0.9356	variável 7		0.4078	0.5237

Fonte: O autor (2021).

Tabela C.3: Variáveis mais significativas usando ANOVA com dados do índice, mercado e *feature engineering* - índice IFIX - parte 1

Variável	**h**	**F Score**	**P-Valor**	**Variável**	**h**	**F Score**	**P-Valor**
variável 1	1	6.2343	0.0139	IPCA mês t-1	3	1.6189	0.2057
variável 6		5.7609	0.0179	variável 4		1.4637	0.2288
IPCA mês t-1		1.7294	0.1910	variável 7		1.2217	0.2713
variável 7		1.1058	0.2951	IPCA mês t-2		0.9798	0.3243
mês t-1		0.9378	0.3348	IPCA mês t-0		0.6015	0.4395
mês t-0		0.8469	0.3593	variável 6		0.5934	0.4427
variável 4		0.6316	0.4283	variável 1		0.2431	0.6229
IPCA mês t-0		0.5627	0.4547	IBOV mês t-0		0.1303	0.7187
variável 2		0.5381	0.4646	IBOV mês t-2		0.1046	0.7470
mês t-2		0.3789	0.5394	variável 3		0.0611	0.8051
IPCA mês t-2		0.1692	0.6815	mês t-0		0.0150	0.9028
IBOV mês t-2		0.1445	0.7045	IBOV mês t-1		0.0076	0.9309
IBOV mês t-0		0.0445	0.8333	mês t-2		0.0062	0.9376
variável 3		0.0438	0.8346	variável 2		0.0025	0.9601
IBOV mês t-1		0.0207	0.8858	mês t-1		0.0001	0.9942

Fonte: O autor (2021).

Tabela C.4: Variáveis mais significativas usando ANOVA com dados do índice, mercado e *feature engineering* - índice IFIX - parte 2

Variável	**h**	**F Score**	**P-Valor**	**Variável**	**h**	**F Score**	**P-Valor**
variável 7		0.6927	0.4070	IPCA mês t-2		4.4174	0.0379
variável 4		0.4228	0.5169	variável 4		3.8782	0.0515
IPCA mês t-2		0.3928	0.5321	IPCA mês t-1		1.2913	0.2583
variável 2		0.3151	0.5757	variável 7		0.9064	0.3432
mês t-2		0.3148	0.5758	variável 6		0.4277	0.5145
mês t-1		0.3130	0.5770	variável 1		0.3994	0.5287
IPCA mês t-1		0.2623	0.6095	mês t-2		0.3503	0.5552
mês t-0	6	0.2396	0.6255	variável 2	12	0.3289	0.5675
variável 1		0.2316	0.6312	mês t-1		0.2849	0.5946
variável 3		0.2300	0.6324	mês t-0		0.1105	0.7403
IBOV mês t-2		0.2275	0.6343	IBOV mês t-2		0.0773	0.7815
IBOV mês t-1		0.2261	0.6353	variável 3		0.0728	0.7878
IBOV mês t-0		0.1132	0.7371	IPCA mês t-0		0.0644	0.8002
variável 6		0.1005	0.7518	IBOV mês t-1		0.0613	0.8048
IPCA mês t-0		0.0261	0.8720	IBOV mês t-0		0.0206	0.8861

Fonte: O autor (2021).

Anexo D - Tabela com dados da composição do fundo imobiliário HGRE11 e valores de alguns indicadores

Tabela D.1: Relação dos imóveis que compõe o fundo HGRE11, com alguns indicadores.

Região	**Nome do Imóvel**	**ABL (m^2)**	**Unidades**	**Participação**	**Vacância**	**Classificação**
Berrini	Ed. Brasilinterpart	887	4	5.76%	30.00%	Lajes Individuais (BB)
Berrini	Roberto Sampaio Ferreira	3.250	6	40.00%	33.33%	Lajes-Part. Relevante (BB)
Berrini	Berrini One	1.0794	15	33.14%	18.39%	Lajes Individuais (AAA)
SCMC[1]	Transatlântico	1.579	3	6.94%	100.00%	Lajes Individuais (BB)
SCMC[1]	Centro Empresarial SP	2.844	2	1.43%	0.00%	Monousuário (BB)
SCMC[1]	BB Antônio Chargas	4.259	2	100.00%	0.00%	Monousuário (BB)
SCMC[1]	Ed. Chucri Zaidan	21.906	1	100.00%	0.00%	Monousuário (AA)
Paulista	Ed. Paulista Star	10.593	16	100.00%	100.00%	Torre Corporativa (A)
Paulista	Torre Martiniano	17.600	22	100.00%	100.00%	Torre Corporativa (A)
Faria Lima	Mario Garnero	3.654.2	5	15.00%	0.00%	Lajes-Part. Relevante (BB)
Faria Lima	Ed. Faria Lima	4.440	10	17.48%	0.00%	Lajes-Part. Relevante (BB)
Zona Norte São Paulo	Totvs	21.100	1	100.00%	0.00%	Monousuário (AAA)
Centro de São Paulo	LIQ Alegria	19.049	1	100.00%	0.00%	Monousuário (C)
RM[2] de São Paulo	Sercom Taboão	16.488	1	100.00%	0.00%	Monousuário

Continua na próxima página

[1]Região que compreende Santo Amaro, Chácara Santo Antônio, Morumbi e Churi

[2]Região Metropolitana

Região	Nome do Imóvel	ABL (m^2)	Unidades	Participação	Vacância	Classificação
Interior de São Paulo	Empresarial Dom Pedro	14.529 (Galpão) 4.496 (Galpão) 6.518 (Galpão)	1	100.00%	13.20%	Outros
Barueri	Ed. Jatobá - CBOP	16.289	1	50.00%	10.37%	Lajes-Part. Relevante (A)
Rio de Janeiro	Torre Rio Sul	1.717	8	1.27%	43.75%	Lajes Individuais (BB)
Rio de Janeiro	Teleporto	2.310	6	5.55%	19.77%	Lajes Individuais (BB)
Curitiba	GVT Curitiba	7.708	1	100.00%	0.00%	Monousuário
Porto Alegre	Guaíba	10.660	1	100.00%	55.00%	Torre Corporativa

Fonte: O autor (2021).

Apêndice - Conceitos Básicos e Preliminares Usados no Trabalho

Neste apêndice apresentamos as definições dos conceitos e termos básicos utilizados neste trabalho. A lista de termos está em ordem alfabética.

Data Frame

São itens de duas dimensões normalmente usados para arquivar base de dados. As duas dimensões de um *data frame* podem ter medidas/unidades diferentes.

Erros do tipo I e tipo II

Os erros do tipo I e do tipo II são os erros que podem ocorrer em testes de hipótese e desejamos que sejam evitados ou que ocorram o mínimo possível. O erro tipo I acontece quando rejeitamos a hipótese nula H_0 quando na realidade ela é verdadeira, conforme Tabela A.1. O erro do tipo II acontece quando aceitamos uma hipótese nula que não possui veracidade.

Tabela A.1: Tipos de erros que podem ocorrer no teste de hipótese.

Realidade	**Decisão do Teste**	
	Aceitamos H_0	Rejeitamos H_0
H_0 verdadeira	$1-\alpha$	erro tipo I α
H_0 falsa	erro tipo II β	$1-\beta$

Fonte: O autor (2021).

Em termos de probabilidade, erro tipo I é $P(Rejeitar\ H_0|H_0\ verdadeira)$ e erro tipo II é $P(Aceitar\ H_0|H_0\ falsa)$.

Feature Engineering

Feature engineering é uma expressão utilizada para designar o conjunto de técnicas aplicadas para a criação de novas variáveis explicativas [13]. Essas técnicas vão desde transformações matemáticas em variáveis já existentes para tirar mais informações até a criação de novas variáveis (*features*) a partir da compreensão da área de negócios.

GridSearch

Técnica usada computacionalmente, para testar e encontrar os melhores hiper-parâmetros [26]. Para utilizar o GridSearch passamos uma lista de parâmetros para

o computador, que testará as combinações possíveis por intermédio da validação cruzada e informará a melhor combinação dos hiperparâmetros.

Hipótese Nula e Hipótese Alternativa

As hipóteses nula e alternativa são as hipóteses que compõem os testes estatísticos chamados por teste de hipótese. A hipótese nula H_0 é a hipótese elaborada com o propósito de ser testada e a hipótese alternativa H_1 é a hipótese contrária a hipótese nula que com base em dados numéricos podem levar a rejeição da hipótese nula [20]. Estas hipóteses são traduzidas em termos matemáticos a partir de suposições pensadas e/ou elaboradas pelo(a) pesquisador(a) que deseja decidir se sua suposição está correta ou não. Por exemplo, uma empresa farmacêutica está desenvolvendo um novo medicamento e tem a intensão de saber se o tempo de efeito desse medicamento é maior que o do concorrente. Para isso, a empresa testa a hipótese de que o tempo médio de efeito é menor que a média do medicamento do concorrente contra a hipótese de que essa suposição é ao contrária. Em termos matemáticos,

$$\begin{cases} H_0 : \mu_{\text{novo medicamento}} = \mu_{\text{medicamento concorrente}}, \\ H_1 : \mu_{\text{novo medicamento}} < \mu_{\text{medicamento concorrente}}. \end{cases}$$

Nível de Significância

O nível de significância é a probabilidade máxima aceitada para erro do tipo I que pode acontecer num teste de hipótese. Normalmente, é denotada por α e o valor comumente usado é 0.05.

Overfitting e Underfitting

Overfitting é o ajuste exagerado do modelo em relação aos dados de treinamento do problema que se esteja resolvendo. Esse sobreajuste resulta em uma medida de avaliação com resultado bom, por exemplo, modelagem que apresenta a acurácia de quase 100% nos dados de treinamento, mas que na validação gera resultados ruins, baixa acurácia.

Underfitting denota o subajuste do modelo, isto é, quando o ajuste do modelo aos dados apresentam resultados que não possui uma performance satisfatória. Desta forma, este modelo apresentará resultados problemáticos e consequentemente terá mais predições de novos dados em classe errada.

Na prática, buscamos um balanceamento nas predições e ajuste dos dados para não haver nem overfitting nem underfitting.

P-valor

O P-valor expressa a probabilidade dos dados terem ocorrido sob à hipótese nula e este valor é calculado a partir de um teste estatístico, que descreve a probabilidade de uma observação estar em consonância a hipótese nula [4]. Para tomada de decisão nos testes de hipóteses, caso o p-valor seja menor ou igual ao nível de significância α rejeitamos a hipótese nula, caso contrário aceitamos a hipótese nula.

Região de Aceitação e Região Crítica

A região de aceitação compreende o intervalo de valores da variável de teste que conduzem a aceitação da hipótese nula H_0 [20]. Caso o valor calculado do teste se

localize nessa região, aceitaremos a hipótese nula. Essa região possui a probabilidade de ocorrer de $1 - \alpha$, em que α é o nível de significância. Essa região fica na área compreendida pelo intervalo

$$\begin{cases} [-valor_{critico},\ valor_{critico}] \text{ no caso bicaudal,} \\ [-valor_{critico},\ \infty) \text{ no caso unilateral à esquerda,} \\ (-\infty,\ valor_{critico}] \text{ no caso unilateral à direita.} \end{cases}$$

A região crítica compreende ao intervalo de valores da variável do teste que direcionam a aceitação da hipótese nula [20]. Caso o valor calculado do teste esteja nesta região, rejeitaremos a hipótese nula. Esta região tem área complementar à área da região de aceitação e possui probabilidade de ocorrência α. O intervalo em que se encontra a região crítica em cada caso é o seguinte

$$\begin{cases} (-\infty,\ -valor_{critico}] \cup [valor_{critico},\ \infty) \text{ no caso bicaudal,} \\ (-\infty,\ -valor_{critico}] \text{ no caso unilateral à esquerda,} \\ [valor_{critico},\ \infty] \text{ no caso unilateral à direita.} \end{cases}$$

Na Figura A.1 observamos as regiões de aceitação e rejeição (crítica) de cada um dos três casos.

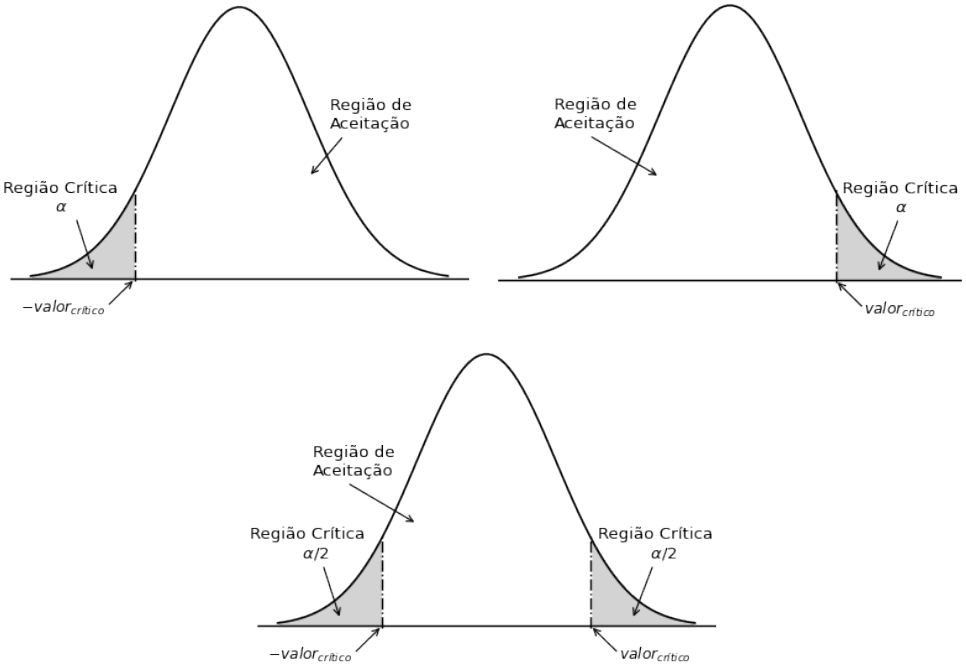

Figura A.1: Opções de testes de hipóteses: unilateral à esquerda, unilateral à direita e bilateral.
Fonte: O autor (2021).

Teste de hipótese

É um método estatístico utilizado para decidir sobre a escolha entre duas hipóteses: a hipótese nula H_0 e a hipótese alternativa H_1. Para essa escolha utilizamos dados reais ou criados computacionalmente baseados em alguma distribuição de probabilidade. Na literatura da área de Estatística, existem vários testes de hipótese,

como teste para média, variância, proporção, etc. E com estes testes buscamos uma tomada de decisão entre rejeitar ou não a hipótese nula.

Teste Unilateral e Teste Bilateral

Em conformidade com a hipótese alternativa H_1, podemos ter os testes de forma unilateral (unicaudal) à esquerda, unilateral à direita ou bilateral (bicaudal). Vejamos na Figura A.1 cada um destas opções de testes de hipótese. Ou seja, testes cujas hipóteses são

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}, \textbf{ ou } \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}, \textbf{ ou } \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases},$$

respetivamente. Nos testes bilaterais o nível de significância α é divido para cada uma das caudas, ou melhor dizendo, cada uma das caudas possui metade da probabilidade do erro do tipo I ($\alpha/2$).

Valor Calculado e Valor Crítico

O valor crítico $valor_{crit}$ é o valor encontrado acerca de uma distribuição de probabilidade que define o limite da região crítica, sendo que nos testes bilaterais temos dois pontos críticos e nos testes unilaterais um ponto crítico.

O valor calculado $valor_{calc}$ é o valor obtido por uma fórmula específica de cada teste de hipótese. Este valor calculado é usado para saber se a hipótese nula será aceita ou rejeitada, para isso comparamos o valor calculado com o valor crítico. Dessa maneira rejeitamos a hipótese nula se o $valor_{calc}$ estiver na região crítica e aceitamos se estiver na região de aceitação.