

**Inspere**  
**Mestrado Profissional em Economia**

Carlos Vinícius Cotrim Fernandes

**Previsão através de técnicas de *machine learning*  
de períodos *risk-off* no Brasil**

**São Paulo**

**2022**

**Carlos Vinícius Cotrim Fernandes**

**Previsão através de técnicas de *machine learning*  
de períodos *risk-off* no Brasil**

Dissertação apresentada ao programa de  
Mestrado Profissional em Economia como  
requisito parcial para obtenção do título de  
Mestre em Economia.

Área de concentração: Economia dos Negócios

Orientador: Prof. Dr. Diogo Abry Guillen

**São Paulo**

**2022**

**Carlos Vinícius Cotrim Fernandes**

**Previsão através de técnicas de *machine learning*  
de períodos *risk-off* no Brasil**

Dissertação apresentada ao programa de  
Mestrado Profissional em Economia como  
requisito parcial para obtenção do título de  
Mestre em Economia.

Orientador: Prof. Dr. Diogo Abry Guillen

**Banca examinadora**

---

Prof. Dr. Diogo Abry Guillen

Inspere

---

Prof. Dr. Miguel Maria Charters de Oliveira Bandeira da Silva

Inspere

---

Prof. Dr. Hellinton Hatsuo Takada

USP

**São Paulo**

**2022**

## Lista de Ilustrações e quadros

<b>Quadro 1 - Variáveis Financeiras.....</b>	<b>16</b>
<b>Quadro 2 - Variáveis Macroeconômicas preditoras.....</b>	<b>17</b>
<b>Figura 1- Algoritmo k-médias com 2 clusters e 5 ativos.....</b>	<b>20</b>
<b>Figura 2- K-médias vs Método Hierárquico.....</b>	<b>21</b>
<b>Quadro 3 - Qualidade da separação.....</b>	<b>22</b>
<b>Figura 3 - Algoritmo k-médias com 2 clusters e 3 ativos.....</b>	<b>23</b>
<b>Figura 4 - Períodos risk-off alinhados com grandes eventos.....</b>	<b>24</b>
<b>Quadro 4 - Matriz de Confusão Teste Random Forest lag 3 meses.....</b>	<b>25</b>
<b>Figura 5 - Contribuição variáveis macroeconômicas lag 3 meses.....</b>	<b>26</b>
<b>Quadro 5 - Matriz de Confusão Teste Random Forest lag 6 meses.....</b>	<b>26</b>
<b>Figura 6 - Contribuição variáveis macroeconômicas lag 6 meses.....</b>	<b>27</b>
<b>Quadro 6 - Matriz de Confusão Teste Random Forest lag 12 meses.....</b>	<b>27</b>
<b>Figura 7 - Contribuição variáveis macroeconômicas lag 12 meses.....</b>	<b>28</b>
<b>Quadro 7 - Overview Performance Preditiva.....</b>	<b>29</b>
<b>Quadro 8 - Análise alocação de portfólios.....</b>	<b>30</b>
<b>Figura 8 - Risco x Retorno – Portfólios.....</b>	<b>30</b>

## Lista de Siglas

**ANFAVEA** – Associação Nacional dos Fabricantes de Veículos

**CART** – *Classification And Regression Trees*

**CAGED** – Cadastro Geral de Empregados e Desempregados

**CDI** – Certificado de Depósito Interbancário

**IPCA** – Índice de preços ao consumidor amplo

**LASSO** – *Least Absolute Shrinkage and Selection Operator*

**VIX** – *Volatility Index*

## **Agradecimentos**

Em um período de tantas incertezas geradas pela COVID-19, gostaria de agradecer as pessoas que de alguma maneira tornaram este trabalho possível.

Inicialmente, à minha família, especialmente, meus pais que sempre me apoiaram com tudo que eu precisava durante a minha vida, além de escutarem as minhas reclamações nos dias difíceis. Ao Insper, que através dos seus docentes e demais funcionários criam um ambiente criativo e amigável ao desenvolvimento do capital humano. Aos meus colegas, em especial: Débora Lopes, Carolina Abreu, Carolina Sverner, Natalia Ishizaki, Rodrigo Arruda e Thales Cunha pelos trabalhos em grupo e compartilhamentos de experiências, mesmo em grande parte remotamente. Ao professor Diogo Guillen, pela orientação sempre atenta e disponível, essencial para o enriquecimento deste trabalho. Por último, à Citrino Gestão de Recursos, pelo apoio moral e financeiro, sobretudo: Carlos Mascarenhas, Guilherme Almeida Prado, Mariana Barros e Fernanda Prado.

## Resumo

Este artigo utiliza técnicas de *machine learning* como K-médias e clusters hierárquicos e o comovimento dos principais ativos financeiros brasileiros ligados aos mercados de juros, câmbio e renda variável para identificar períodos popularmente conhecidos como *risk-on* e *risk-off* no Brasil.

Após identificar os períodos, busco prevê-los com o uso de diversas variáveis macroeconômicas em defasagens de três, seis e doze meses com relação aos períodos encontrados e *Random Forests*. Aqui temos a contribuição relevante de variáveis não frequentemente utilizadas para previsões como dados de emprego e índices de preços.

Por fim, realizo simulações de carteiras com os ativos brasileiros e as previsões obtidas, buscando mostrar uma aplicação possível para os períodos *risk-off* encontrados e as previsões das *Random Forests*, nessas simulações temos ganho de rentabilidade independente da defasagem em relação ao portfólio 60/40 e ganho de rentabilidade das defasagens 6 e 12 meses em relação ao CDI.

Palavras-chave: *Risk-on*, *Risk-off*, K-médias, Cluster Hierárquico, *Random Forest*, *Machine learning*, portfólio 60/40

## **Abstract**

This article uses machine learning techniques such as K-means and hierarchical clusters and the movement of the main Brazilian financial assets to identify periods popularly known as risk-on and risk-off in Brazil.

After identifying the periods, I seek to predict them using several macroeconomic variables in lags of three, six and twelve months in relation to the periods found above and Random Forests. Here we have the relevant contribution of variables not frequently used for forecasts such as employment data and price indices.

Finally, I carry out portfolio simulations with Brazilian assets and the forecasts obtained, seeking to show one possible application for the risk-off periods found and the predictions of Random Forests, in these simulations we have a gain in profitability regardless of the lag in relation to the portfolio 60/ 40 and profitability gain from lags of six and twelve months in relation to the CDI.

**Keywords:** Risk-on, Risk-off, K-means, Hierarchical Cluster, Random Forest, Machine learning, 60/40 portfolio

## Sumário Executivo

Os termos *risk-on* e *risk-off* são frequentemente utilizados pela mídia e especialistas para explicar movimentações nos ativos financeiros em todo o mundo, a cada crise ou euforia nos mercados, eles ressurgem. Na prática, ambos possuem relação com o medo dos investidores, o período *risk-off* é caracterizado quando os investidores se tornam em média mais avessos ao risco, muitas vezes até ignorando os fundamentos das empresas ou dos países e devido a isso, começam a vender os ativos considerados de maior risco, no qual se enquadram o Ibovespa, o real, juros brasileiros e ativos de países emergentes em geral, e buscam ativos mais seguros como títulos do tesouro americano, o CDI e o dólar. Enquanto, em períodos *risk-on* ocorre fenômeno inverso, no qual a aversão ao risco na média diminui e os investidores buscam os ativos de maior risco, perfil em que se enquadram os ativos brasileiros e de outros emergentes ligados aos mercados de renda variável, câmbio e juros.

Diante disso, identificar de forma quantitativa e prever tais períodos se tornou o principal objetivo deste trabalho, pois na literatura até então, os períodos são basicamente classificados através do *Volatility Index* (VIX), uma medida da volatilidade do mercado de ações americano que é amplamente utilizada como *proxy* de aversão ao risco dos investidores, os autores se alternam na forma como identificam essas variações, mas sempre utilizam o índice.

Entretanto, principalmente no caso brasileiro, que tem períodos de aversão sem grande relação com o VIX, devido a crises locais, que não influenciam ou sofrem influência do mesmo, se tornou necessário uma abordagem diferente, assim utilizamos o comovimento dos principais ativos financeiros brasileiros ligados aos mercados de juros, câmbio e bolsa de valores para identificar os períodos de maior aversão ao risco através das técnicas de *machine learning*: k-médias e agrupamentos hierárquicos, os resultados trazem episódios *risk-off* que são facilmente identificáveis quando comparados aos principais fatos macroeconômicos do período.

Assim, passamos a utilizar outras técnicas de *machine learning* e diversas variáveis macroeconômicas com disponibilidade mensal nas defasagens de três, seis

e doze meses para prever tais períodos *risk-off* anteriormente encontrados. Aqui a melhor técnica das utilizadas foi a *Random Forest* em todas as métricas analisadas.

Finalmente, já com as previsões das *Random Forest* nas três diferentes defasagens, realizamos simulações de carteiras com ativos brasileiros em busca de uma aplicação prática da metodologia em gestão de portfólios. O resultado encontrado mostrou que com uma estratégia de investimentos simples, melhora-se consideravelmente o retorno e as medidas de risco com relação ao tradicional portfólio 60/40 adaptado ao Brasil e apenas o retorno em relação ao CDI, evidenciando o valor de se identificar e prever tais períodos.

## Sumário

1. Introdução .....	12
2. Revisão Bibliográfica .....	14
3. Base de Dados .....	16
4. Metodologia .....	18
4.1. Identificação dos períodos risk-off.....	18
4.1.1. Método Hierárquico .....	18
4.1.2. K-médias.....	19
4.1.3. Qualidade da partição .....	22
4.2. Previsões através das variáveis macroeconômicas e Random Forest.....	24
5. Resultados das previsões com variáveis macro .....	25
6. Carteiras de investimento padrão.....	29
7. Conclusão .....	31

## 1. Introdução

A primeira pergunta que deve ser respondida quando discutimos *risk-on* e *risk-off* é sobre o seu significado. Em busca desta resposta, Smales (2016) mostra que “tomados em conjunto, os retornos do mercado financeiro estão intimamente relacionados aos níveis predominantes de medo do investidor”, assim também segundo ele: “*Risk-on* e *risk-off* são termos frequentemente usados por especialistas de mercado e pela mídia, em um esforço para explicar os movimentos de preços nos mercados financeiros. Essencialmente, o comentário atribui as flutuações do mercado às mudanças no nível de aversão ao risco do investidor”.

Ambos os termos possuem relação com o medo dos investidores, o período *risk-off* é caracterizado quando os investidores se tornam em média mais avessos ao risco, muitas vezes até ignorando os fundamentos das empresas ou dos países e devido a isso, começam a vender os ativos considerados de maior risco, no qual se enquadram o Ibovespa, o real, juros brasileiros e ativos de países emergentes em geral, e buscam ativos mais seguros como títulos do tesouro americano, CDI e o dólar. Enquanto, em períodos *risk-on* ocorre fenômeno inverso, no qual a aversão ao risco na média diminui e os investidores buscam os ativos de maior risco, perfil em que se enquadram os ativos brasileiros e de outros emergentes ligados aos mercados de renda variável, câmbio e juros.

Para Smales (2016), De Bock e Carvalho Filho (2015) e McCauley (2013) a melhor forma de identificar os períodos *risk-on* e *risk-off* seria através das variações no *Volatility Index* (VIX), uma medida da volatilidade do mercado de ações americano que é amplamente utilizada como *proxy* de aversão ao risco dos investidores, os autores se alternam na forma como identificam essas variações como, por exemplo, De Bock e Carvalho Filho (2015) que optou por definir o início de um período *risk-off* como o período quando o VIX é 10 pontos percentuais maior do que sua média móvel retroativa de 60 dias. Já Smales (2016) utiliza o VIX diretamente como *proxy* do medo dos investidores e relaciona períodos de VIX alto ou baixo em relação à média com os retornos dos ativos financeiros. Em uma outra forma de abordagem, McCauley (2013) faz um vínculo entre VIX alto e baixo, e os fluxos de capital internacionais, assim em períodos de *risk-on*, os fluxos deixariam os países desenvolvidos e

buscariam os emergentes afetando positivamente os retornos dos ativos financeiros desses países e nos períodos *risk-off* ocorreria o contrário.

Como observamos acima, temos uma lacuna na literatura a respeito de quando ocorre os períodos *risk-off*, ainda mais no Brasil, que tem períodos de aversão sem grande relação com o VIX que é o principal instrumento de identificação utilizado pelos autores citados, assim ao longo deste trabalho, realizamos uma abordagem diferente em que utiliza-se o comovimento dos principais ativos financeiros brasileiros ligados aos mercados de juros, câmbio e bolsa de valores para identificar os períodos de maior aversão ao risco através das técnicas de *machine learning*: k-médias e agrupamentos hierárquicos, os resultados trazem episódios *risk-off* que são facilmente identificáveis, cabe ressaltar que ao analisar o caso brasileiro o VIX acaba tendo uma performance menos eficiente que apenas a utilização dos retornos dos ativos brasileiros, provavelmente porque durante o período estudado tivemos algumas crises locais, que não influenciam ou sofrem influência do VIX. Assim, o resultado obtido é uma forma quantitativa para identificar os períodos *risk-off* brasileiros com ganhos sobre a utilização do VIX.

Agora no tocante às previsões desses períodos encontrados com as variáveis macroeconômicas brasileiras mensais disponíveis no período, a técnica *Random Forest* foi superior aos outros métodos avaliados em todos os *lags* avaliados (3 meses, 6 meses e 12 meses à frente), possibilitando assim a utilização dos períodos *risk-off* encontrados para formação de carteira de investimentos padrão que superam o portfólio 60/40 adaptado em todas as defasagens estudadas e nos *lags* de 6 e 12 meses superam também o CDI do período, evidenciando outra possível aplicação para o método.

Além da introdução, o artigo se estrutura em outras seis seções. A seção 2 descreve brevemente o que já foi estudado na literatura expondo similaridades e diferenças; a seção 3 mostra os dados utilizados; a seção 4 a metodologia aplicada tanto na identificação dos períodos como nas previsões; a seção 5 expõe os resultados das previsões; a seção 6 comenta sobre a performance de carteiras de investimento padrão; e, por fim, a seção 7 traz a conclusão do estudo, ressaltando as contribuições do artigo.

## 2. Revisão Bibliográfica

A discussão sobre os períodos *risk-on* e *risk-off* ainda é recente, a identificação dos períodos tem em comum o VIX como indicador chave, o VIX é um índice calculado através das opções das maiores empresas do mercado americano (S&P 500) e expressa uma visão consensual sobre a volatilidade futura esperada pelo mercado para os próximos 30 dias, após sua criação em 1993, com o passar dos anos se tornou comum a utilização do indicador como uma *proxy* de aversão ao risco.

Por conta disso, McCauley (2013) fez um vínculo entre VIX alto e baixo, e os fluxos de capital internacionais, assim em períodos de *risk-on* os fluxos deixariam os países desenvolvidos e buscariam os emergentes afetando positivamente os retornos dos ativos financeiros desses países e nos períodos *risk-off* ocorreria o inverso.

Destacou também, que o efeito que ocorre nos investimentos entre os períodos *risk-on* e *risk-off* é assimétrico, pois os *carry trades* alavancados, utilizam moedas de *low-yield* para investir em moedas de *high-yield* e segundo McCauley(2013): “quando em *risk-on*, essas posições produzem uma série de ganhos pequenos; quando em *risk-off*, eles podem produzir grandes perdas repentinas”.

De Bock e Carvalho Filho (2015) analisaram o comportamento das moedas durante os episódios de *risk-off* e para identificar os períodos de maneira quantitativa definiram o início de um período *risk-off* como o período em que o VIX é 10 pontos percentuais maior do que sua média móvel retroativa de 60 dias, alegaram que as vantagens do VIX são principalmente sua alta frequência e sua conformidade com os eventos reconhecidamente relevantes para o mercado nas últimas duas décadas, como por exemplo: o atentado em 11 de setembro, a falência do Lehman e a crise na Grécia, a crítica a esse método é que ele acaba sendo arbitrário.

Os autores enfatizaram que os riscos presentes nas economias do G-3 estão conduzindo à maior frequência de episódios de *risk-off*, principalmente, após 2007, porque os investidores estariam mais preocupados com os riscos de cauda para estabilidade financeira global advindos da Zona do Euro e dos Estados Unidos.

Por fim, Smales (2016) estuda a relação entre as mudanças no nível de medo do investidor (medido pelo VIX) e os retornos do mercado financeiro, ele encontra que à medida que o VIX aumenta, acontece uma queda nos mercados de ações, nos

rendimentos de títulos e valorização do dólar, evidenciando uma busca por segurança quando o medo aumenta, que é refletida nos comportamentos dos ativos financeiros nas três maiores classes: renda variável, juros e câmbio. O autor não busca identificar períodos *risk-on* e *risk-off*, seu objetivo foi mostrar o efeito do VIX nos outros ativos financeiros e sua forte relação com o medo do investidor, entretanto relata sobre os termos: “*Risk-on* e *risk-off* são termos comumente usados por comentaristas de mercado e pela mídia, em um esforço para explicar os movimentos de preços nos mercados financeiros. Essencialmente, o comentário atribui as flutuações do mercado às mudanças no nível de aversão ao risco do investidor”.

Em busca de aperfeiçoar as formas de identificação, esse artigo propõe um método de classificação quantitativo sem a utilização do VIX, baseado no agrupamento através do comovimento dos principais ativos financeiros brasileiros utilizando técnicas de *machine learning*, durante a análise de agrupamentos, que é uma forma de análise exploratória de dados em que as observações são divididas em diferentes grupos que compartilham características comuns, que é justamente o que se busca com a classificação dos períodos *risk-on* e *risk-off*, utilizei o algoritmo do Hartigan and Wong (1979) que faz a partição dos dados em um número  $k$ , pré determinado de grupos, de tal maneira que a soma dos quadrados dos pontos aos centros dos grupos é minimizada, o principal problema do algoritmo anterior é a escolha arbitrária do número  $k$  de grupos que é de certa forma mitigada pela utilização do método do agrupamento hierárquico que procede de maneira sequencial juntando a cada etapa os dois grupos mais próximos, seguindo até que sobre apenas um, fornecendo assim sem estipular antecipadamente o número de grupos os resultados mais eficientes. Existem diversos critérios de se formar os grupos hierárquicos, o critério de Murtagh e Legendre (2014), também conhecido como critério de Ward que minimiza a variância total dentro de cada grupo e encontra os grupos que levam ao aumento mínimo na variância total dentro do grupo após a fusão foi o que apresentou os resultados mais satisfatórios.

Já para as previsões dos períodos *risk-on* e *risk-off* identificados, respaldei-me em Spyridon D.Vrontos, Galakis e Ioannis D. Vrontos (2021) que aplicaram:

“Uma ampla variedade de técnicas de *machine learning* para previsão de recessão nos EUA. Especificamente, os métodos empregados incluem técnicas de regularização, como Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) e Elastic Net, classificadores de Análise

Discriminante, classificadores Bayesianos e árvores de classificação e regressão (CART), como *Bagging*, *Random Forests* e *Boosting*. Além disso, os modelos de benchmark Probit e Logit que empregam a curva de juros como um preditor de recessão são incluídos na análise para comparação com a literatura existente.”

No decorrer deste artigo, aplico a metodologia empregada pelos autores na utilização de *lags* (3 meses ,6 meses e 12 meses), além do benchmark com Logit e CART. Entretanto, meu estudo se resume apenas a técnica das *Random Forest* (Breiman, 2001) que constitui uma técnica de *machine learning* que normalmente gera excelente desempenho preditivo em comparação com a maioria dos modelos conhecidos.

Por fim, crio carteiras de investimento padrão com os resultados das previsões e comparo com dois *benchmarks*, o primeiro o CDI e o segundo o portfólio 60/40 que é uma estratégia tradicional defendida por Ambachtsheer (1987), que demonstra que a diversificação entre ações e títulos, apesar de simples gera ganhos para as carteiras no longo prazo.

### 3. Base de Dados

Esta seção visa apresentar os dados utilizados e fazer uma breve explanação sobre suas características. Foram utilizados retornos mensais com valores do fim do mês ano sobre ano anterior do período entre abril de 2001 até junho de 2021 dos seguintes ativos financeiros retirados da plataforma Bloomberg:

#### Quadro 1: Variáveis Financeiras

Variáveis Financeiras	Ticker Bloomberg	Unidade	Transformação
Ibovespa	IBOV Index	(%)	ano sobre ano
dólar-real	USDBRL Currency	(%)	ano sobre ano
Juros 3 anos	BZAD3Y Index	(%)	ano sobre ano
Juros 1 anos	BZAD1Y Index	(%)	ano sobre ano
inclinação de 3 meses com 3 anos	BZAD3Y Index - BZAD3M Index	(%)	ano sobre ano

Fonte: Elaboração própria

yCabe ressaltar que a escolha da transformação ano contra ano anterior sobre os retornos mensais, apesar de gerar um aumento da autocorrelação, resultou em uma persistência desejável para classificação dos grupos *risk-on* e *risk-off*, grupos que alternam com muita frequência em pouco espaço de tempo, estariam reagindo puramente aos movimentos dos preços de mercado e não a mudanças reais de aversão ao risco dos investidores.

Outro ponto de destaque é o vencimento fixo adotado para os juros, esse método é frequentemente utilizado para análise desta classe de ativo, já que sem o vencimento fixo, não estaríamos comparando a mesma coisa ao longo do tempo.

As variáveis macroeconômicas foram selecionadas de acordo com a disponibilidade mensal e dentro do período já citado acima, sendo utilizadas com defasagens de 3, 6 e 12 meses para as previsões:

#### **Quadro 2: Variáveis Macroeconômicas preditoras**

<b>Variável Macroeconômica Preditora</b>	<b>Unidade</b>	<b>Transformação</b>
Fenabreve: Venda de Veículos - Fenabreve	(%)	ano sobre ano
SCPC (em número absoluto de consultas)	(%)	ano sobre ano
Serasa: Indicador de Atividade do Comércio	(%)	ano sobre ano
ANFAVEA: Produção de automóveis	(%)	ano sobre ano
ONS: Carga de energia Total MW médio	(%)	ano sobre ano
IBS: Produção Aço Total	(%)	ano sobre ano
RFB: IPI TOTAL em milhões	(%)	ano sobre ano
ABCR: Fluxo de Veículos Pesados	(%)	ano sobre ano
CAGED: Saldo de vagas	nº absoluto	12 meses acumulado
IBGE: IPCA	(%)	ano sobre ano
RFB: Arrecadação	(%)	ano sobre ano
FGV: IGP-M	(%)	ano sobre ano
FGV: IPA produtos agrícolas	(%)	ano sobre ano
FGV: IPA produtos industriais	(%)	ano sobre ano

Fonte: Elaboração própria

A maioria dos dados macroeconômicos foram coletados através do sistema de séries temporais do Banco Central do Brasil, mas alguns deles foram retirados nos sites das respectivas instituições fornecedoras da informação.

## 4. Metodologia

### 4.1. Identificação dos períodos *risk-off*

A análise de agrupamentos é uma forma de análise exploratória de dados em que as observações são divididas em diferentes grupos que compartilham características comuns.

Como já exposto durante a revisão da literatura, foi encontrado por McCauley (2013), De Bock e Carvalho Filho (2015) e Smales (2016) uma relação entre o comportamento dos ativos financeiros, VIX e os períodos *risk-on* e *risk-off*.

Por isso, a fim de identificar os períodos *risk-on* e *risk-off* brasileiros, iremos agrupar os mais importantes ativos financeiros brasileiros ligados aos juros, câmbio e renda variável com o objetivo de construir grupos (*clusters*), garantindo a seguinte propriedade: dentro de um grupo, as observações devem ser tão semelhantes quanto possível, enquanto as observações pertencentes a grupos diferentes devem ser tão diferentes quanto possível, assim quanto mais distantes os grupos formados, mais útil se torna a análise.

Existem dois tipos principais de agrupamento: k-médias e o método hierárquico. O primeiro é geralmente usado quando o número de classes é fixado com antecedência, enquanto o segundo é usualmente empregado para um número desconhecido de classes e ajuda a determinar esse número ótimo para ser utilizado no k-médias inclusive.

A escolha do número de *clusters* tem sempre uma subjetividade e depende muito dos objetivos da análise, recordando que o nosso objetivo principal é prever períodos conhecidos pelos agentes do mercado e mídia como *risk-on* e *risk-off* no mercado financeiro brasileiro.

#### 4.1.1. Método Hierárquico

Iniciarei a análise através dos clusters hierárquicos, pois eles permitem determinar o número ideal de clusters sem necessidade de escolha *a priori*. O

funcionamento do *cluster* hierárquico é bem intuitivo, o algoritmo começa colocando cada ponto em seu próprio *cluster*, então em seguida, mescla os 2 pontos que estão mais próximos um do outro com base na matriz de similaridades entre os *clusters*. A consequência é que há um *cluster* a menos. Ele então recalcula as distâncias entre os *clusters* novos e antigos e os salva em uma nova matriz de similaridades que será usada na próxima etapa.

Finalmente, as etapas 1 e 2 são repetidas até que todos os *clusters* sejam mesclados em um único *cluster* incluindo todos os pontos, cabe destacar que é um processo sequencial e que foi utilizada a média dos retornos ano contra ano dos ativos.

Existem diversos métodos para medir a distância entre os *clusters*, conhecidos como métodos de ligação, eu adotei o critério de Ward que minimiza a variância total dentro do *cluster* e encontra o par de *clusters* que leva ao aumento mínimo na variância total dentro do *cluster* após a fusão.

Aplicando o método no *software* R versão 4.0.5 através do pacote *hclust* sobre os retornos mensais ano contra ano do período entre abril de 2001 até junho 2021 dos seguintes ativos financeiros: índice Ibovespa, dólar-real, juros 3 anos, juros 1 ano e inclinação de 3 meses com 3 anos temos pela análise do dendrograma, diagrama que mostra a hierarquia e relação dos agrupamentos, que o número ideal de *clusters* é dois.

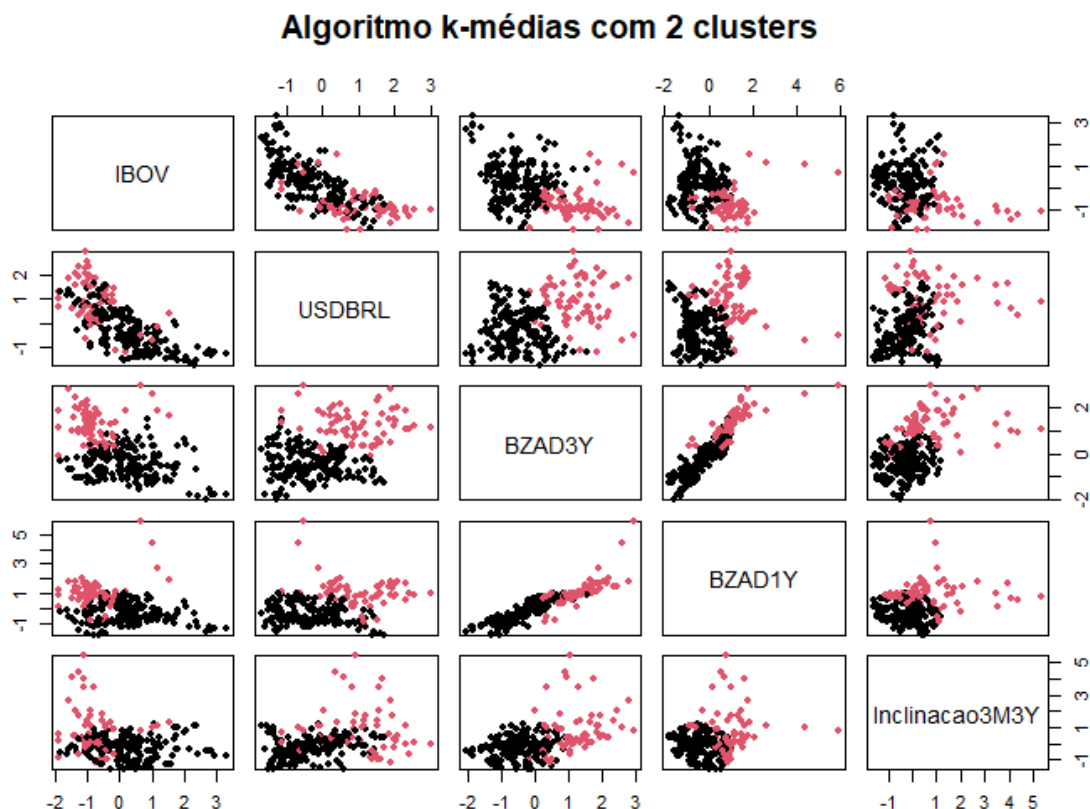
#### **4.1.2.K-médias**

O K-médias, que foi encontrado através do pacote *kmeans* do *software* R versão 4.0.5, assume que a medida de dissimilaridade é a distância entre os retornos mensais ano contra ano dos ativos financeiros selecionados e possui como ideia principal, definir k centroides, um para cada cluster pré-determinado. Esses centroides devem ser colocados de forma astuciosa, segundo a qual a melhor opção é colocá-los o mais longe possível um do outro. A próxima etapa é pegar cada ponto pertencente a um determinado conjunto de dados e associá-lo ao centroide mais próximo. Quando nenhum ponto está pendente, a primeira etapa é concluída e um

grupo inicial é formado. Neste ponto, é necessário recalcular  $k$  novos centroides como centros dos clusters resultantes da etapa anterior. Após esses  $k$  novos centroides, uma nova ligação deve ser feita entre os mesmos pontos de dados e o novo centroide mais próximo. O resultado desse processo recorrente, pode-se notar que os  $k$  centroides mudam sua localização passo a passo até que nenhuma mudança seja feita, a recorrência leva os centroides a convergência em um centroide estável, que embora não forneça garantias de que é a melhor solução encontrada, gera bons resultados.

Assim, dado os resultados dos clusters hierárquicos temos  $k$  igual a dois como o ideal:

**Figura 1: Algoritmo k-médias com 2 clusters e 5 ativos**

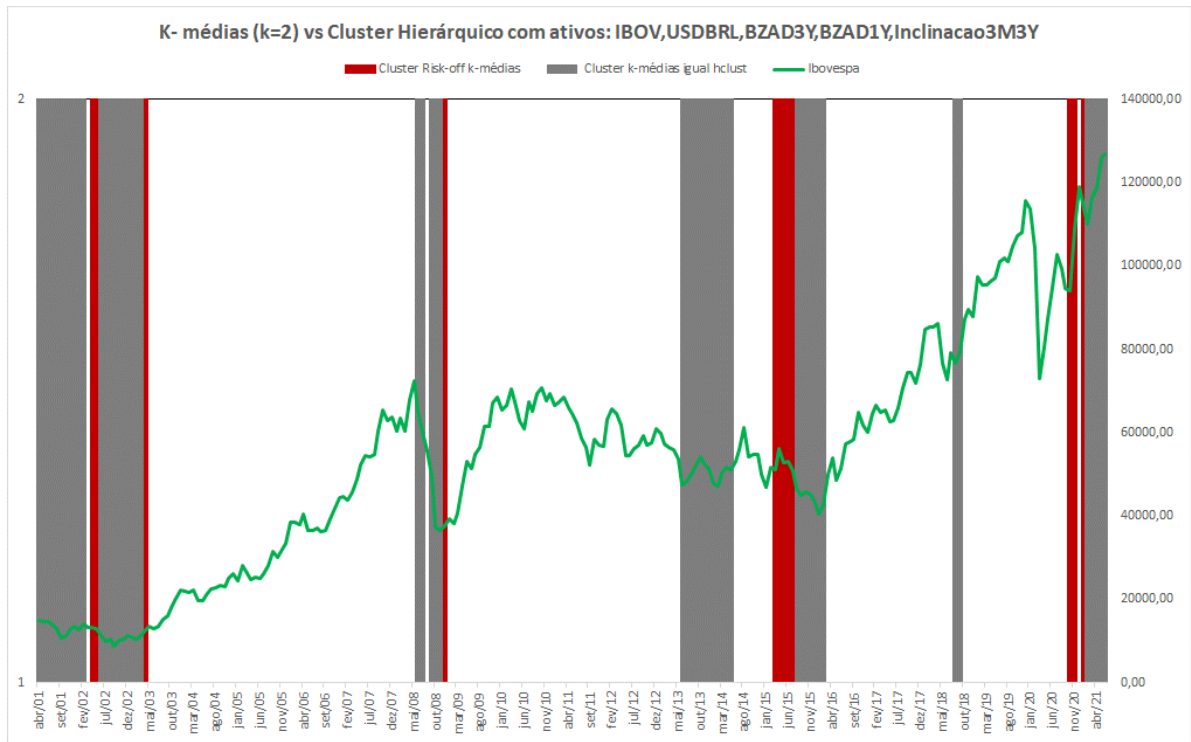


Fonte: Elaboração própria

Observa-se na figura 1, os dois clusters formados pelo k-médias, fica claro que a divisão é razoável em grande parte dos ativos com exceção para a inclinação. Cabe ressaltar, que ao observar os grupos formados pelo k-médias e método hierárquico

existe uma grande semelhança dos resultados com apenas 16 das 243 observações sendo classificadas de forma diferente, observações em vermelho na figura 2.

**Figura 2: K-médias vs Método Hierárquico**



Fonte: Elaboração própria

Assim, não existe diferença relevante em usar o método hierárquico e k-médias, vou seguir a análise neste trabalho apenas com o k-médias, pois facilitará a avaliação da seleção das melhores variáveis financeiras selecionadas para a melhor partição dos dois grupos: *risk-on* e *risk-off*.

### 4.1.3. Qualidade da partição

A qualidade de uma partição k-médias é encontrada calculando a porcentagem do TSS “explicado” pela partição usando a seguinte fórmula:

$$\frac{BSS}{TSS} \times 100\%$$

Onde BSS e TSS representam *Between Sum of Squares* e *Total Sum of Squares*, respectivamente. Quanto maior a porcentagem, melhor será a pontuação e, conseqüentemente, a qualidade da divisão.

Nota-se que a classificação com mais classes sempre será a melhor, pois a contribuição do BSS será maior.

Por outro lado, o “modelo” será mais complexo e talvez o agrupamento deixe de ter interesse. No caso extremo em que  $k = n$  (cada observação é uma classe única), temos  $BSS = TSS$ , mas a partição perdeu toda a função de existir.

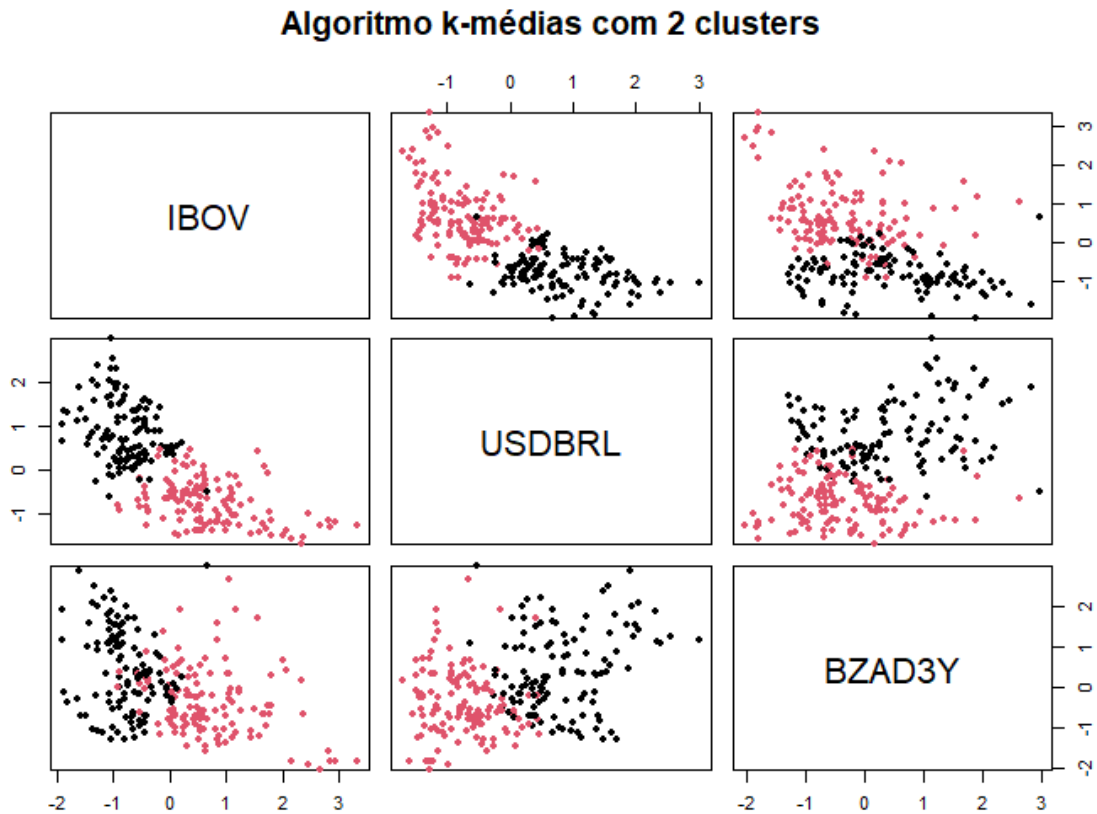
A fim de encontrar a melhor seleção de ativos financeiros para separação dos dois grupos formamos o quadro 3, na qual o melhor resultado encontrado foi na separação com apenas três ativos: Ibovespa, dólar-real e taxa de juros de 3 anos. Nota-se que mesmo com 3 *clusters* o resultado dos melhores ativos é o mesmo.

#### Quadro 3: Qualidade da separação

Qualidade da separação (%) número de clusters	Ativos financeiros brasileiros		
	IBOV,USDBRL,BZAD3Y,BZAD1Y,Inclinacao3M3Y	IBOV,USDBRL,BZAD3Y,Inclinacao3M3Y	IBOV,USDBRL,BZAD3Y
2	34,96	36,43	43,96
3	49,41	51,45	58,30

Fonte: Elaboração própria

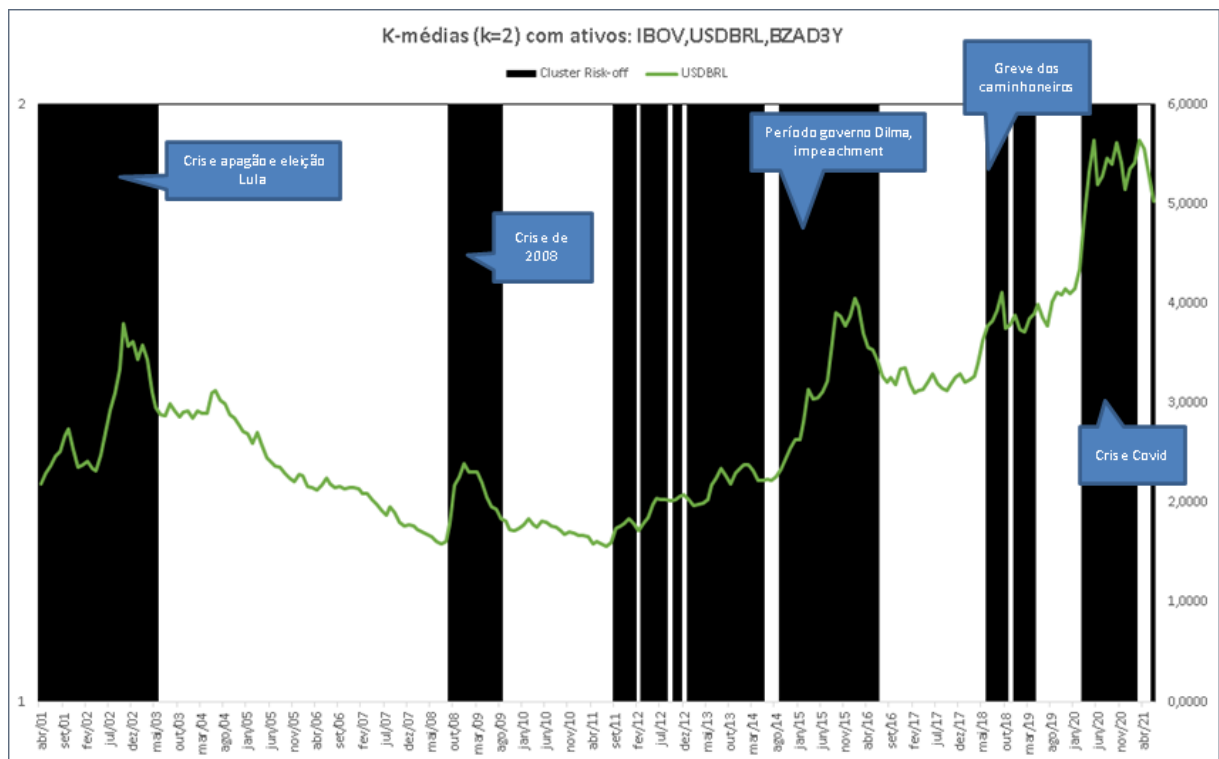
Figura 3: Algoritmo k-médias com 2 clusters e 3 ativos



Fonte: Elaboração própria

O agrupamento com três ativos gerou períodos *risk-off* alinhados com os grandes eventos políticos e macroeconômicos do país no período como podemos ver na figura 4, atestando empiricamente a qualidade da partição gerada.

**Figura 4: Períodos risk-off alinhados com grandes eventos**



Fonte: Elaboração própria

#### **4.2. Previsões através das variáveis macroeconômicas e *Random Forest***

Após a formação da partição em dois grupos através do comovimento dos principais ativos financeiros brasileiros, vamos a escolha das variáveis macroeconômicas para previsão, essa acabou levando em conta a disponibilidade de dados para o período analisado que foi entre abril de 2001 até junho de 2021 e sua frequência mensal, com esse critério muitas variáveis mais conhecidas acabaram não sendo mensais, como o PIB por exemplo ou não estando condizente com o período estudado caso da produção industrial, vendas no varejo e IBC-BR. É importante o período começar em 2001, para abranger a crise energética e eleitoral do Brasil que ocorreu na época, com isso foi preferível perder essas variáveis que encurtar o período.

A *Random Forest*, criada por Breiman (2001), constitui uma técnica alternativa de aprendizado de máquina que normalmente gera maior precisão/desempenho preditivo em comparação com modelos de árvore de classificação e regressão padrão.

As *Random Forests* podem reduzir a variância do classificador agregando um número de árvores com baixa correlação, neste artigo utilizamos o *default* do pacote *RandomForest* do *software* R que é 500 árvores, obtidas usando diferentes amostras de treinamento *Bootstrap*.

## 5. Resultados das previsões com variáveis macro

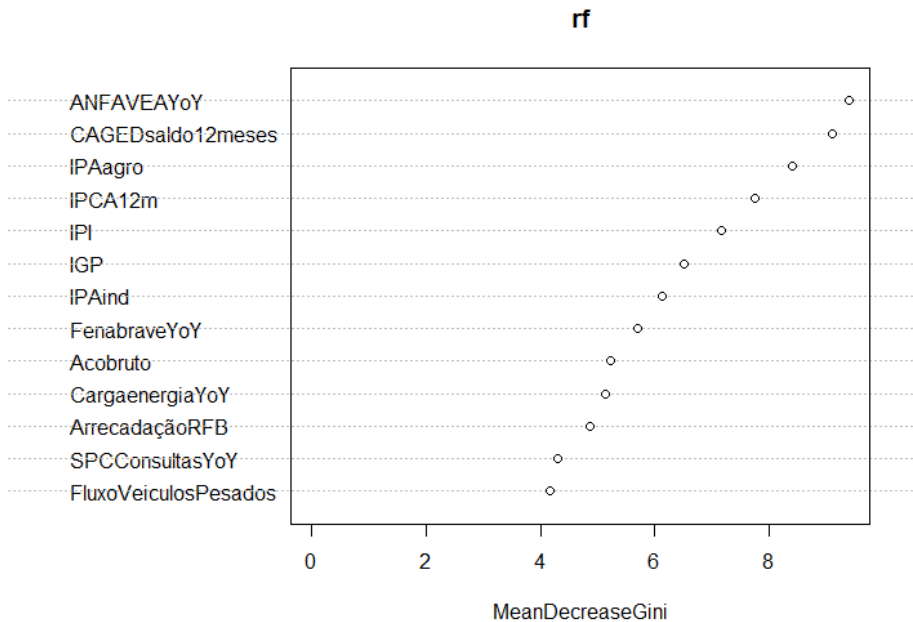
Através das variáveis macroeconômicas do quadro 2 com defasagem de 3 meses em relação aos grupos *risk-on* e *risk-off* e da amostra de um total de 243 observações separadas em 170 ou 70% no grupo treinamento e 73 ou 30% no grupo teste para validação cruzada, além do pacote *RandomForest* no *software* R temos a seguinte matriz de confusão no grupo de teste:

**Quadro 4: Matriz de Confusão Teste *Random Forest lag 3 meses***

		Matriz de Confusão RF lag 3 meses	
		Classificação observada	
Classificação Prevista		Risk-on	Risk-off
Risk-on		34	7
Risk-off		4	28

Fonte: Elaboração própria

**Figura 5: Contribuição variáveis macroeconômicas lag 3 meses**



Fonte: Elaboração própria

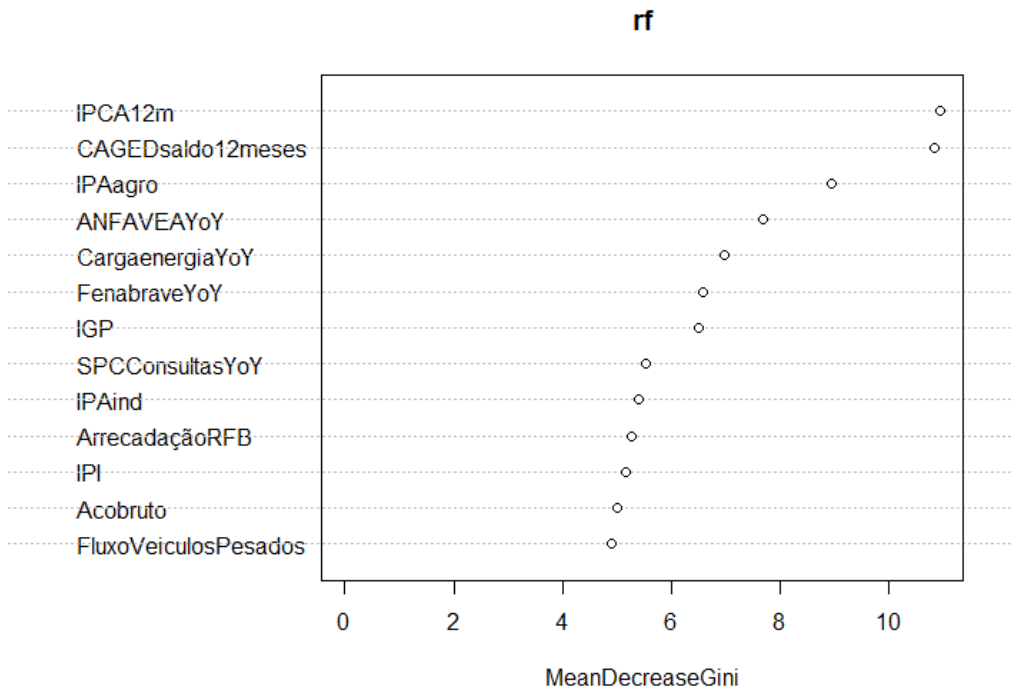
Na figura 5 temos uma tentativa de desmistificar a caixa preta do algoritmo, mostrando as variáveis com “maior importância”. Vamos repetir o processo abaixo com as outras defasagens:

**Quadro 5: Matriz de Confusão Teste *Random Forest* lag 6 meses**

Matriz de Confusão RF lag 6 meses		
Classificação observada		
Classificação Prevista	Risk-on	Risk-off
Risk-on	29	10
Risk-off	1	21

Fonte: Elaboração própria

**Figura 6: Contribuição variáveis macroeconômicas lag 6 meses**



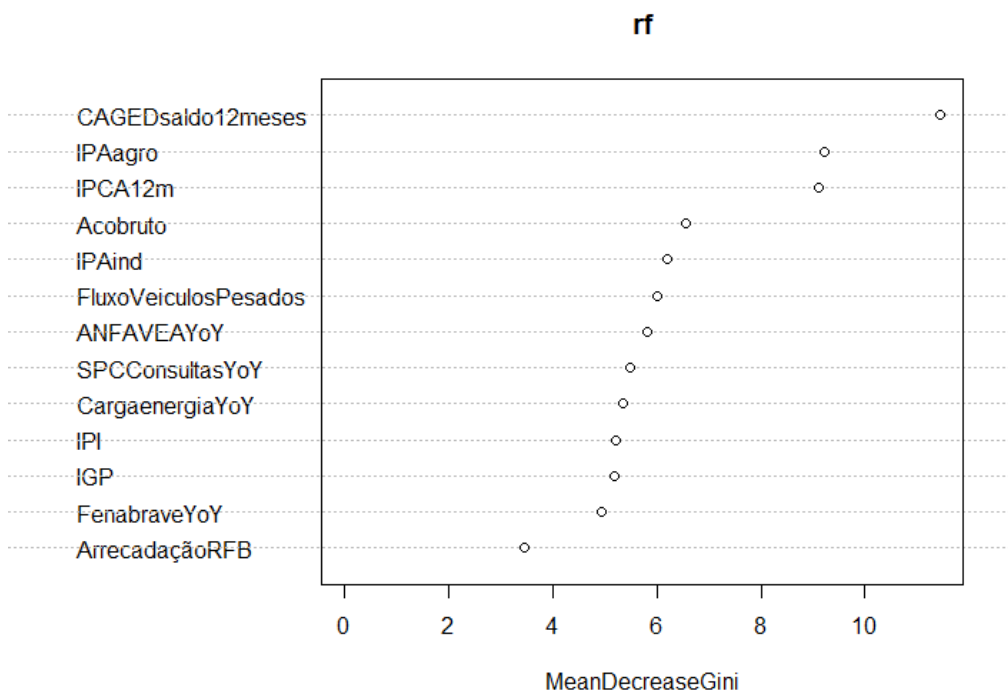
Fonte: Elaboração própria

**Quadro 6: Matriz de Confusão Teste *Random Forest* lag 12 meses**

Matriz de Confusão RF lag 12 meses		
Classificação observada		
Classificação Prevista	Risk-on	Risk-off
Risk-on	31	3
Risk-off	7	32

Fonte: Elaboração própria

**Figura 7: Contribuição variáveis macroeconômicas lag 12 meses**



**Fonte: Elaboração própria**

Nota-se que nas figuras 5,6 e 7 os índices de preços, principalmente o índice de preços ao consumidor amplo (IPCA) 12 meses tem contribuição relevante independente da defasagem, juntamente com o saldo do Cadastro Geral de Empregados e Desempregados (CAGED) 12 meses que também apresentou um comportamento semelhante. Já a produção de veículos da Associação Nacional dos Fabricantes de Veículos (ANFAVEA) só teve destaque relevante na defasagem mais curta, figura 5.

Todas as matrizes de confusão, quadro 4,5 e 6 apresentaram bons resultados preditivos, abaixo no quadro 7, veremos um resumo com algumas métricas para analisar a qualidade das previsões.

## Quadro 7: Overview Performance Preditiva

Overview Performance Preditiva					
lag 3 meses variáveis macroeconômicas					
Método	AUC	Acurácia	Sensibilidade	Especificidade	Precisão balanceada
Logit (Threshold = 0.5)	0,7940	0,2877	0,2667	0,3023	0,2845
Single Tree	0,8414	0,7534	0,7500	0,7576	0,7538
Random Forest	0,9274	0,8493	0,8293	0,8750	0,8521
lag 6 meses variáveis macroeconômicas					
Método	AUC	Acurácia	Sensibilidade	Especificidade	Precisão balanceada
Logit (Threshold = 0.5)	0,7656	0,3279	0,28	0,3611	0,3206
Single Tree	0,6478	0,5574	0,5429	0,5769	0,5599
Random Forest	0,8785	0,8197	0,7436	0,9545	0,8491
lag 12 meses variáveis macroeconômicas					
Método	AUC	Acurácia	Sensibilidade	Especificidade	Precisão balanceada
Logit (Threshold = 0.5)	0,7301	0,3425	0,3438	0,3415	0,3426
Single Tree	0,8139	0,7534	0,7500	0,7576	0,7538
Random Forest	0,9098	0,8630	0,9118	0,8205	0,8661

Fonte: Elaboração própria

No quadro 7, observamos a performance superior da *Random Forest* quando comparada com a *single tree* e ao Logit em todos os *lags* estudados e em todas as medidas de qualidade analisadas. Na área embaixo da curva (AUC), métrica bastante utilizada em aprendizado de máquina, o resultado supera 0,9 em duas das três defasagens, mostrando resultados bem satisfatórios. Além disso o modelo é equilibrado tendo sensibilidade, meses classificados como *risk-off* sobre total de meses *risk-off* e especificidade, meses classificados como *risk-on* sobre os meses classificados como *risk-on* parecidos, assim a precisão balanceada mostra resultado semelhante a área debaixo da curva (AUC). Dado isso fica evidente a superioridade do método em comparação ao *Logit* e *Single tree* independente da métrica observada.

## 6. Carteiras de investimento padrão

Dada a boa performance observada, vamos avaliar se a formação dos períodos *risk-off* e a previsão dos mesmos tem capacidade de gerar retornos superiores ao CDI e a alocação de portfólios clássica 60/40 adaptada para o Brasil, na qual emprego a alocação de 60% em Ibovespa e 40% no título pré-fixado de *duration* fixa de 3 anos.

No quadro 8, temos que o maior retorno total foi o da carteira ideal que utiliza a classificação encontrada pelo k-médias dos períodos *risk-off* e aplica em CDI nesses períodos, mantendo alocação 60/40 adaptada nos períodos *risk-on*, nas demais carteiras utilizo a classificação de *risk-off* encontrada pelas *Random Forest* em seus *lags*: 3, 6 e 12 meses e realizo a mesma estratégia de investimentos descrita acima, agora para os períodos previstos.

## Quadro 8: Análise alocação de portfólios

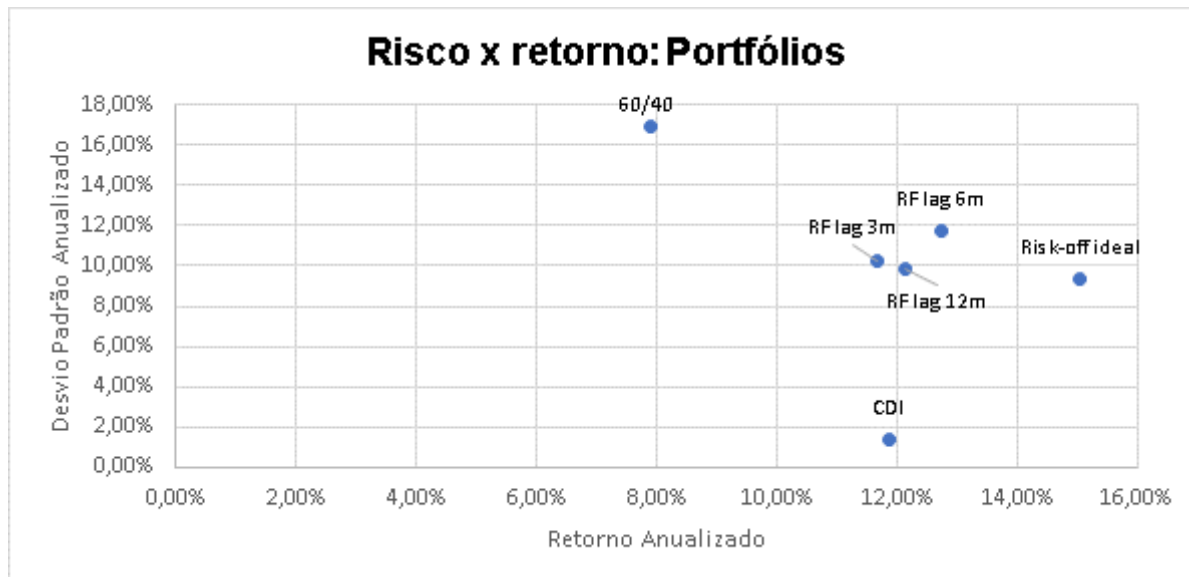
Análise alocação de Portfólios						
Método	Retorno total	Retorno anualizado	Desvio Padrão Anualizado	Maior queda mensal	Máximo Drawdown	Sharpe
CDI	875,55%	11,85%	1,36%	-	-	-
Portfólio 60/40	370,12%	7,91%	16,94%	-18,21%	-37,07%	-0,25
Portfólio Risk-off ideal	1627,72%	15,04%	9,37%	-8,17%	-12,01%	0,40
Portfólio Risk-off lag 3 meses	842,78%	11,67%	10,28%	-17,24%	-28,64%	-0,02
Portfólio Risk-off lag 6 meses	1045,94%	12,74%	11,77%	-18,21%	-28,64%	0,09
Portfólio Risk-off lag 12 meses	925,08%	12,13%	9,87%	-8,17%	-15,67%	0,03

Fonte: Elaboração própria

O segundo maior retorno absoluto é o dos períodos encontrados através da defasagem de 6 meses das variáveis macroeconômicas, seguido pelo lag de 12 meses, já o lag de 3 meses foi o único que ficou abaixo do CDI.

Cabe ressaltar também, o desempenho do portfólio 60/40 adaptado ao Brasil, que apresentou o pior retorno, maior volatilidade e maior *drawdown* entre os portfólios analisados, assim ao comparar as estratégias Random Forest, todas apresentaram ganhos significativos em relação ao clássico portfólio 60/40, independente da defasagem e das métricas observadas no quadro 8.

Figura 8: Risco x Retorno – Portfólios



Fonte: Elaboração própria

Ao analisar, a performance através do risco versus retorno das carteiras e compará-las com o CDI, apesar do retorno total maior na carteira ideal e nas defasagens 6 e 12 meses, o ganho de rentabilidade não compensou o aumento de risco e o CDI seria a melhor escolha.

Observamos pela figura 8 e quadro 8, que na ótica do risco a melhor das carteiras formadas pelas *random forest* é a de defasagem de 12 meses que apresentou menor máximo *drawdown* e menor desvio padrão dentre as defasagens.

Uma alternativa que poderia trazer ganhos em relação ao CDI também na ótica do risco seria modificar a alocação para os períodos *risk-on*, já que o portfólio 60/40 adaptado ao Brasil não teve grande performance e a utilização de CDI nos períodos *risk-off* trouxe contribuição bastante positiva em relação a estratégia pura, evidenciando assim que está parte da estratégia foi acertada.

## 7. Conclusão

A partição dos períodos em *risk-on* e *risk-off* realizada pelo k-médias com suporte do método hierárquico teve resultados bastante satisfatórios ao confrontarmos com o ocorrido, pois praticamente todos os períodos com evidente aversão ao risco no Brasil como eleições de 2002, grande crise financeira de 2008, impeachment da Dilma, greve dos caminhoneiros e a crise do covid de 2020 foram considerados corretamente como períodos *risk-off*, assim o artigo trouxe uma forma quantitativa através do comovimento dos ativos financeiros brasileiros para identificar os períodos historicamente reconhecidos como *risk-off*.

Além da partição, temos testada a capacidade de indicadores macroeconômicos não tão convencionais para a previsão dos períodos *risk-off*, o que pode orientar a ampliação de outras pesquisas que envolvem previsão através de indicadores macroeconômicos, sendo nesse artigo os dados de maior destaque: saldo 12 meses do CAGED, IPCA 12 meses e alguns outros índices de preços.

Destaque ainda ao próprio método de classificação através das técnicas de *machine learning*, principalmente *Random Forest* que não é muito comum na literatura e trouxe excelentes resultados em comparação a clássica alocação 60/40, podendo assim ser uma opção para gestão de portfólios, mostrando possível aplicação prática.

## Referências

- AMBACHTSHEER, K. P. Pension fund asset allocation: In defense of a 60/40 equity/debt asset mix. **Financial analysts journal**, v. 43, n. 5, p. 14–24, 1987.
- BEKAERT, G.; HOEROVA, M.; LO DUCA, M. Risk, uncertainty and monetary policy. **Journal of monetary economics**, v. 60, n. 7, p. 771–788, 2013.
- BREIMAN, L. Bagging predictors. **Machine learning**, v. 24, n. 2, p. 123–140, 1996.
- BREIMAN, L. Random Forest. **Machine learning**, v. 45, p. 5–32, 2001.
- BRUNNERMEIER, M. K.; NAGEL, S.; PEDERSEN, L. H. Carry Trades and Currency Crashes. **NBER macroeconomics annual**, v. 23, n. 1, p. 313–348, 2008.
- DE BOCK, R.; DE CARVALHO FILHO, I. The behavior of currencies during risk-off episodes. **Journal of international money and finance**, v. 53, p. 218–234, 2015.
- FERNANDES, M.; MEDEIROS, M. C.; SCHARTH, M. Modeling and predicting the CBOE market volatility index. **Journal of banking & finance**, v. 40, p. 1–10, 2014.
- FLEMING, J.; OSTDIEK, B.; WHALEY, R. E. Predicting stock market volatility: A new measure. **Journal of futures markets**, v. 15, n. 3, p. 265–302, 1995.
- GIOT, P. Relationships between implied volatility indexes and stock index returns. **The Journal of Portfolio Management**, v. 31, n. 3, p. 92–100, 2005.
- GRAHAM, M.; NIKKINEN, J.; PELTOMÄKI, J. Web-based investor fear gauge and stock market volatility: An emerging market perspective. **Journal of Emerging Market Finance**, v. 19, n. 2, p. 127–153, 2020.
- HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-means clustering algorithm. **Journal of the Royal Statistical Society. Series C, Applied statistics**, v. 28, n. 1, p. 100, 1979.
- LLOYD, S. Least squares quantization in PCM. **IEEE transactions on information theory**, v. 28, n. 2, p. 129–137, 1982.
- MCCAULEY, R. N. Risk-on/risk-off, capital flows, leverage, and safe assets. **SSRN Electronic Journal**, 2013.
- MURTAGH, F.; LEGENDRE, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? **Journal of classification**, v. 31, n. 3, p. 274–295, 2014.
- SMALES, L. A. Risk-on/Risk-off: Financial market response to investor fear. **Finance research letters**, v. 17, p. 125–134, 2016.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society. Series B, Statistical methodology**, v. 63, n. 2, p. 411–423, 2001.

VRONTOS, S. D.; GALAKIS, J.; VRONTOS, I. D. Modeling and predicting U.S. recessions using machine learning techniques. **International journal of forecasting**, v. 37, n. 2, p. 647–671, 2021.