

Monitoramento do mercado de ativos brasileiro: uma proposta de pipeline de dados para detecção de bolhas financeiras

Uiliam B. Bomfim¹, Flávia Maristela S. Nascimento²

¹ Análise e Desenvolvimento de Sistemas – Instituto Federal da Bahia (IFBA) – Salvador – BA – Brasil

² Análise e Desenvolvimento de Sistemas – Instituto Federal da Bahia (IFBA) – Salvador – BA – Brasil

wf3@outlook.com.br, flaviamsn@ifba.edu.br

Resumo. *Especulações e crises financeiras podem causar problemas econômicos à população, especialmente em países emergentes. Por isso, bancos centrais, reguladores e formuladores de políticas buscam identificar bolhas financeiras antecipadamente, para que possam tomar medidas que proporcionem estabilidade econômica. Neste sentido, os algoritmos de aprendizado de máquina (machine learning) podem ser utilizados para dar apoio ao processo de tomada de decisão. De fato, procedimentos desta natureza podem alertar usando estratégias de vigilância, conduzidas por bancos centrais e reguladores fiscais, além dos agentes econômicos (famílias, empresas e governo) que buscam resguardar seus interesses. Porém, a utilização desses procedimentos se revela uma tarefa desafiadora, devido à necessidade de conhecimentos específicos, principalmente nas áreas de programação e estatística, o que pode ser um desafio para alguns analistas que não tenham expertise nestas áreas. Para mitigar esta lacuna, este trabalho apresenta um pipeline de dados que automatiza o processo de ingestão, transformação, carga dos dados e machine learning, além de disponibilizar um painel com gráficos e métricas que facilitam a identificação de bolhas financeiras das empresas listadas na B3 (Brasil, Bolsa e Balcão). Utilizou-se para atingir este objetivo o procedimento PSY, um algoritmo de machine learning útil para detecção de bolhas e crises, e a ferramenta Databricks para criar o pipeline de dados. A análise dos resultados revela dois períodos de crescimento do Dividend Yield e Dividend JSCP Yield, coincidindo com eventos macroeconômicos, evidenciando a capacidade do pipeline de dados de identificação de bolhas na B3. Apesar dos resultados positivos, o estudo destaca desafios, como a influência do contexto macroeconômico nas décadas de 90 e 2000 e uma quantidade limitada de empresas listadas na B3, ressaltando a importância contínua da evolução do pipeline para fortalecer sua utilidade na detecção de bolhas financeiras e contribuir para a estabilidade econômica do país.*

1. Introdução

Comportamentos especulativos e crises financeiras podem causar grandes problemas à economia. Por isso, bancos centrais, reguladores e legisladores buscam identificar com antecedência os sinais de alerta para tomar medidas para estabilidade financeira [Phillips e Shi 2020].

Do ponto de vista econômico, uma bolha é um desvio do valor fundamental [Phillips and Shi 2020]. O valor fundamental de um ativo é uma medida que representa seu valor intrínseco, determinado por meio de uma análise aprofundada dos aspectos financeiros e econômicos associados a ele, incluindo demonstrativos financeiros, perspectivas de crescimento, dividendos, taxas de juros e riscos, independentemente do preço de mercado atual [Bragagnolo 2020, Chaim e Laurini 2019].

Identificar o valor fundamental de alguns ativos não é uma tarefa trivial. Para Phillips et al. (2015a,b), uma bolha pode ser definida como um comportamento explosivo de preços. Evidências históricas indicam que as bolhas são tipicamente transitórias, com fases alternadas de expansão e colapso nos preços dos ativos. Durante a fase de expansão de uma bolha financeira, uma consequência típica é a má alocação de recursos, uma vez que os fundos são direcionados para a especulação de ativos e não para empresas produtivas [Shi e Phillips 2022].

A partir de meados dos anos 90, uma série de crises financeiras vem impactando a economia mundial. A principal diferença entre as crises que tiveram início na década de 90 e as anteriores é que as "novas" crises não se restringem apenas ao país de origem e seus vizinhos, mas sim, reverberam mundialmente [Dias Júnior 2022]. Desde a década de 1990 até a primeira década do século XXI, ocorreram diversas turbulências internacionais. Assim, os episódios de crises financeiras e políticas do período estudado foram Crise da Ásia (1997-1998), Crise da Rússia (1998), Crise Brasileira (1999), Bolha da Internet (2000), Crise Argentina (2001), Ataque Terrorista de 11 de Setembro (2001), Crise Brasileira (2002) e Crise do Subprime (2007-2008), Início da crise das commodities (2010), Greve dos caminhoneiros (2018) e Covid-19(2020) [Bourgard e Gomes 2017, Deloitte 2020].

Apesar de características próprias de cada crise financeira, existe um consenso que as crises decorrentes de especulação financeira causam sérios danos à economia [Phillips e Shi 2020]. Por isto, identificar o capital especulativo e o risco de crédito é de fundamental importância para os órgãos reguladores e agentes econômicos. Os Bancos Centrais têm enfrentado debates sobre como lidar com a possível formação de bolhas de ativos. Após a crise financeira de 2008, a atuação dos Bancos Centrais migrou para uma posição que preconiza uma ação preventiva contra bolhas. [Espindola 2015]. Apesar da necessidade crescente de identificar bolhas financeiras com antecedência, existe ainda uma gama limitada de ferramentas para detecção. Por isto, no meio acadêmico, a mera existência de bolhas, e a capacidade de as detectar, permanece em debate.

Na literatura recente, observa-se um esforço considerável na identificação de bolhas, destacando-se o papel proeminente dos algoritmos de *machine learning*. Essa área de estudo concentra-se na construção de modelos computacionais capazes de aprender e tomar decisões autônomas com base em dados, possibilitando a integração dessas técnicas para a detecção de bolhas financeiras [Kufel et al. 2023]. Uma inovação importante é a aplicação de testes recursivos de raiz unitária de cauda direita, conhecidos por sua eficácia em identificar bolhas [Monschang e Wilfling 2020]. Esses testes são frequentemente usados em monitoramento de bolhas, como por exemplo, sup-ADF-style, CUSUM(cumulative sum control chart) e PWY. Essas técnicas têm

aplicações em vários mercados, incluindo ações, commodities e imóveis [Caspi e Graham 2017].

Na busca por identificação de bolhas, os estudos têm se voltado para a aplicação de algoritmos de *machine learning*. Contudo, constata-se uma lacuna na disponibilidade de ferramentas completas (ou *scripts*) que possam conduzir o processo integral de pipeline de dados. Esse processo envolve a transferência e transformação de dados provenientes de diversas fontes até um destino específico, com o propósito de gerar *insights* ou análises de negócios [Densmore 2021]. A utilização de pipelines de dados possibilita o tratamento abrangente dos dados, desde a sua ingestão até a etapa de visualização, tornando mais acessível a informação para os usuários finais.

Trabalhos recentes sobre mecanismos de detecção econométrica mostraram a eficácia de procedimentos recursivos na identificação e datação de bolhas financeiras [Phillips e Shi, 2020]. Nesse contexto, o procedimento PSY é amplamente utilizado como um diagnóstico de alerta precoce de comportamento semelhante a bolhas [Hu e Oxley 2018]. Monschang e Wilfling (2020), demonstraram que o procedimento PSY possui um nível de assertividade de início e fim das bolhas quando comparado a outras técnicas.

Dada a importância para os agentes do mercado financeiro, reguladores e o impacto que bolhas especulativas podem causar na economia real, justifica-se criar um pipeline de dados automatizado para identificação de bolhas financeiras.

Este trabalho apresenta um pipeline de dados para automatizar o processo de ingestão, transformação, carga dos dados e *machine learning*, além de disponibilizar um painel com gráficos e métricas que facilitam a identificação de bolhas financeiras. Utilizou-se para atingir este objetivo o procedimento PSY, um algoritmo de *machine learning* útil para detecção de bolhas e crises, e a ferramenta Databricks para criar o pipeline de dados.

2. Estado da Arte

Existem bastante desafios relacionados à melhoria de detecção prévia de bolhas e crises financeiras. Esta seção analisa diferentes estudos encontrados na literatura sobre mecanismos de detecção econométrica de bolhas e a eficácia destes procedimentos na identificação e datação de bolhas financeiras.

A proposta de Phillips et al. (2015a) observou que a utilização de dados de longos períodos históricos apresenta um desafio econométrico mais sério devido à complexidade da estrutura não linear e dos mecanismos de quebra que são inerentes aos fenômenos de bolhas múltiplas. Para enfrentar esse desafio, eles desenvolveram um novo método recursivo de janela flexível que é mais adequado para este tipo de dado. O método é uma versão generalizada do teste *sup augmented* Dickey-Fuller (ADF) e oferece uma estratégia consistente de carimbo de data (*timestamp*) para a origem e término de múltiplas bolhas. Simulações mostram que o teste melhora significativamente o poder discriminatório e leva a ganhos de poder distintos quando ocorrem múltiplas bolhas.

Escobari et al. (2017) buscam identificar períodos de bolha financeira nos principais mercados acionários da América Latina. Foram utilizados os métodos recursivos baseados em *Augmented* Dickey-Fuller recentemente desenvolvidos. O autor concluiu que, dependendo das bolhas no S&P 500, índice ponderado de valor de mercado que representa as 500 maiores empresas dos Estados Unidos, existem fortes ligações entre os episódios de bolha nos mercados acionários da América Latina. Além disso, os períodos de bolha financeira na América Latina começaram mais cedo e duraram mais do que os períodos de bolha nos Estados Unidos durante a crise financeira de 2008.

Os testes recursivos baseados em *Augmented* Dickey-Fuller tornaram-se uma ferramenta popular para testar a existência de bolhas nos preços das ações. Estes testes exigem dados contínuos sobre a distribuição de dividendos que nem sempre estão disponíveis em alguns mercados. Caspi e Graham (2017) demonstraram que é possível contornar este problema aplicando o teste a uma bolha de ações utilizando o índice *book-to-market*. Os resultados demonstraram a robustez frente a presença potencial de volatilidade não estacionária, ou seja, quando as propriedades estatísticas de uma série temporal não permanecem constantes ao longo do tempo.

Hu e Oxley (2018) concentraram-se em testar bolhas nos mercados de ações e imobiliário do Japão do primeiro trimestre de 1970 ao quarto trimestre de 1999, usando o teste de Phillips et al. (2015a) e os métodos econométricos de Greenaway-McGrevy e Phillips (2016) para explorar a possibilidade de contágio entre estes dois mercados. Os autores ofereceram evidências significativas de bolhas em ambos os mercados durante este período no Japão. Estas descobertas podem ajudar a compreender porque a bolha imobiliária do Japão entrou em colapso após a bolha dos preços das ações, à medida que o comportamento semelhante a uma bolha do mercado de ações migra para o mercado imobiliário.

Chaim e Laurini (2019), exploraram a narrativa da bolha Bitcoin, analisando a volatilidade dos preços diários e de alta frequência em comparação com ativos financeiros tradicionais (S&P500, EUR-USD, ouro e petróleo). A análise foi realizada entre o período de 2013 e 2017, utilizando o estimador não paramétrico de Florens-Zmirou. Observou-se que o Bitcoin é significativamente mais volátil, apresentando dinâmicas de variação de nível distintas. A aplicação do modelo de volatilidade estocástica revela uma bolha no Bitcoin durante a subamostra de janeiro de 2013 a abril de 2014.

Hu (2023) aplicou a abordagem PSY com o novo procedimento *bootstrap* à famosa British Railway Mania da década de 1840. Os resultados dos testes fornecem evidências de comportamento explosivo nos preços das ações das ferrovias em 1835/1836 e 1846, que estão relacionados ao *boom* ferroviário em 1836 e à mais proeminente Railway Mania em meados da década de 1840, respectivamente.

Philips e Shi (2020) forneceram uma implementação, usando a linguagem R, da popular estratégia de monitoramento proposta por Phillips et al. (2015a,b), juntamente com um novo procedimento de *bootstrap* projetado para mitigar o impacto potencial da heterocedasticidade em algoritmos de teste recursivos. A heterocedasticidade refere-se à variabilidade não constante dos erros em um modelo estatístico, e a introdução desse

novo procedimento visa lidar de maneira mais robusta com tais variações. Esta metodologia tem-se mostrado eficaz na detecção de bolhas e crises [Phillips e Shi, 2017; Phillips et al., 2015a,b] e é agora utilizada por investigadores acadêmicos, economistas de bancos centrais e reguladores fiscais. Esses aplicativos são implementados usando o pacote *psymonitor* R [Phillips et al., 2018].

Utilizando como ponto de partida a implementação em R desenvolvida por Phillips e Shi (2020), propomos a criação de um *pipeline* de dados que automatiza o processo de ingestão, transformação, carga dos dados e *machine learning*, além de disponibilizar um painel com gráficos e métricas que facilitam a identificação de bolhas financeiras.

Quadro 1 - Estudos recentes relacionados a identificação de bolhas

Autor(es) e Ano de Publicação	Aspectos Investigados	Principais Evidências
Phillips et al. (2015a)	Desenvolvimento de método recursivo para detecção de bolhas financeiras em dados históricos extensos.	Introduziu um método recursivo de janela flexível baseado no teste ADF, melhorando o poder discriminatório e identificando múltiplas bolhas.
Escobari (2017)	Identificação de períodos de bolha nos mercados acionários da América Latina e análise das ligações entre bolhas na América Latina e S&P 500.	Identificou fortes ligações entre bolhas nos mercados latino-americanos e eventos no S&P 500. Bolhas na América Latina começam mais cedo e duram mais que nos EUA em 2008.
Caspi e Graham (2017)	Contornar a falta de dados contínuos em testes de bolhas, aplicando o teste a uma bolha de ações usando o índice <i>book-to-market</i> .	Ofereceu uma solução para a limitação de dados contínuos, revelando a eficácia do teste utilizando o índice <i>book-to-market</i> e sua robustez diante da presença potencial de volatilidade não estacionária.
Hu e Oxley (2018)	Concentraram-se em testar bolhas nos mercados de ações e imobiliário do Japão do primeiro trimestre de 1970 ao quarto trimestre de 1999.	Contribuiu para a compreensão da relação entre bolhas nos mercados de ações e imobiliário no Japão, destacando a presença significativa de bolhas em ambos os setores durante o período analisado.

Chaim e Laurini (2019)	Explora a narrativa da bolha Bitcoin, analisando a volatilidade em comparação com ativos tradicionais.	Destacou a significativa volatilidade do Bitcoin em comparação com ativos tradicionais, utilizando o estimador não paramétrico. Identificou uma bolha específica no Bitcoin durante o período de janeiro de 2013 a abril de 2014.
Hu (2023)	Aplica a abordagem PSY com novo procedimento <i>bootstrap</i> à British Railway Mania da década de 1840. Evidências de comportamento explosivo nos preços das ações das ferrovias em 1835/1836 e 1846, relacionados ao boom ferroviário e Railway Mania.	Demonstrou o comportamento explosivo nos preços das ações de ferrovias relacionado ao boom ferroviário em 1836 e à Railway Mania em meados da década de 1840.
Phillips e Shi (2020)	Fornecimento de implementação em R da estratégia de monitoramento proposta por Phillips et al. (2015a,b), com procedimento de <i>bootstrap</i> .	Contribuiu para a disseminação da estratégia de monitoramento, tornando-se amplamente utilizada. O novo procedimento de <i>bootstrap</i> mitigou impactos potenciais da heterocedasticidade, garantindo eficácia na detecção de bolhas e crises.

3. Proposta

3.1 Arquitetura da Solução

Neste tópico descreveremos a solução proposta. A Figura 1 descreve o pipeline de dados que será construído para este trabalho. Um pipeline de dados é uma sequência de fases de processamento que geralmente envolve ingestão, transformação, carga e *machine learning* e visualização de dados. Nesta seção detalharemos este processo.

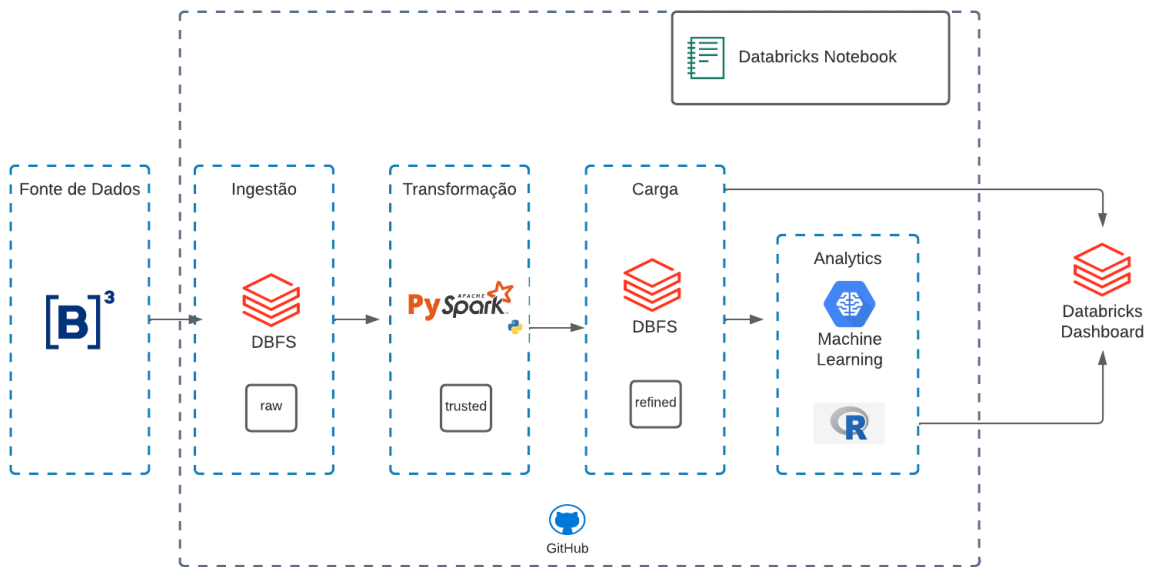


Figura 1: Fluxo de dados

1. Ingestão: Primeira etapa no processo de inserção de dados no ambiente de dados. Neste trabalho realizamos a ingestão de dados a partir do site da B3 (Brasil, Bolsa e Balcão), utilizando a técnica de *web scraping*. Este processo, apresentado na Figura 2, permite ler e compreender os dados que serão utilizados nos processos seguintes.

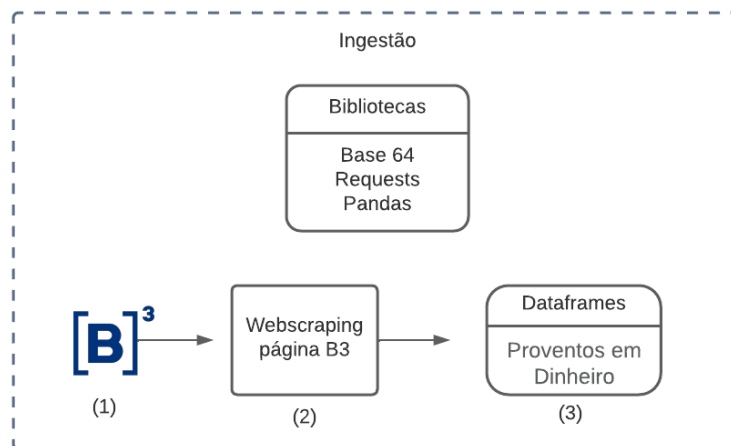


Figura 2: Ingestão de Dados: (1) Fonte de dados da B3(*Brasil, Bolsa e Balcão*) (2) Raspagem de dados do site da B3 (3) Dados brutos que consideram informações de distribuição de dividendos das empresas.

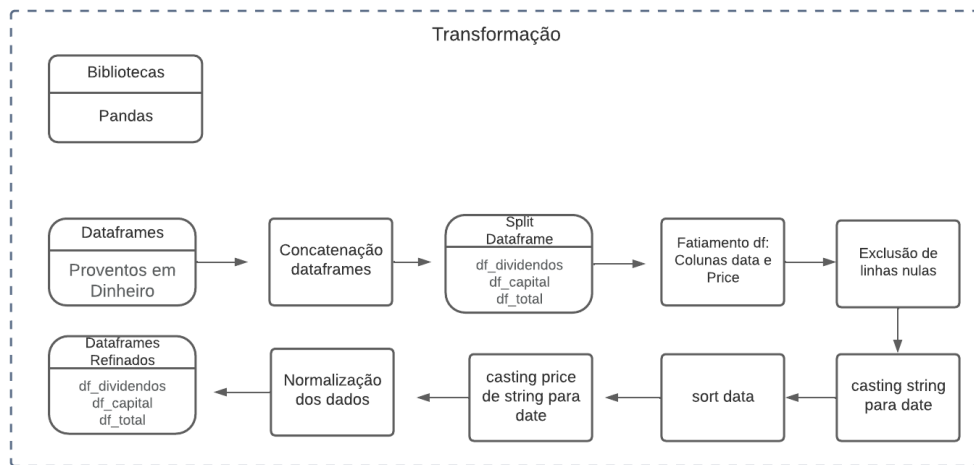


Figura 3: Transformação de dados

2. **Transformação:** Nesta etapa, os dados brutos provenientes da fase de ingestão são submetidos a um processo de transformação, visando criar uma estrutura adequada para consulta e consumo pelos usuários. Em nosso caso, a transformação envolveu a conversão dos dados para *dataframes*¹ pandas. As etapas compreenderam a concatenação dos *dataframes* brutos, a realização de limpeza para remoção de valores nulos, a seleção das colunas de data e preço, a conversão de tipos de dados, e a subdivisão dos *dataframes* em três partes distintas: um *dataframe* de dividendos, um *dataframe* de juros sobre o capital próprio, e um *dataframe* combinando ambas as categorias (Ver Figura 3).

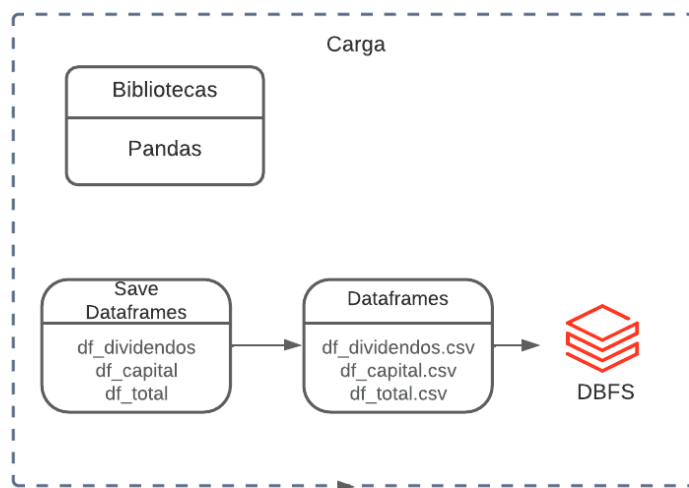


Figura 4: Carga

¹ O termo *dataframe* é utilizado aqui como uma estrutura de dados análoga a uma tabela de dados bidimensional

3. Carga: Esta fase caracteriza-se por disponibilizar dados de qualidade para os usuários. Os *dataframes* resultantes da fase de transformação foram carregados no DBFS (*Databricks File System*) no formato csv, conforme apresentado na Figura 4

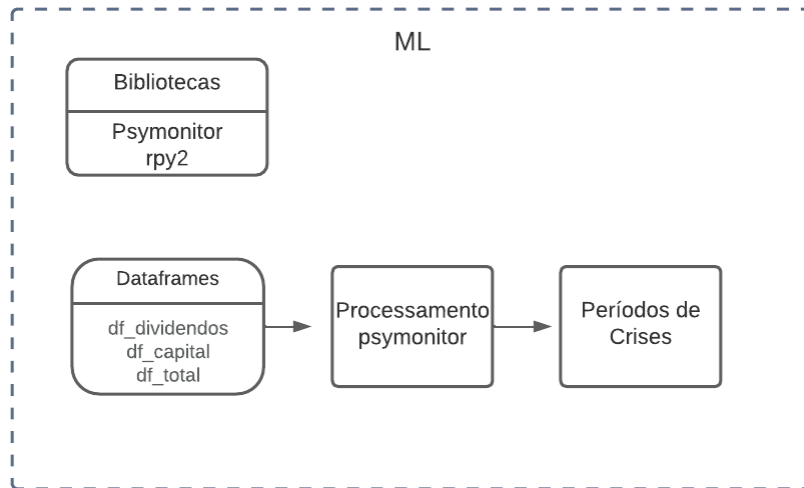


Figura 5: *Machine Learning*

4. Machine Learning: Fase caracterizada pelo uso de um algoritmo para extrair padrões a partir dos dados disponíveis e utilizá-los para fazer previsões [Usmani e Shamsi 2021]. Neste trabalho utilizamos o algoritmo chamado procedimento PSY. Este procedimento, apresentado na Figura 5, é amplamente utilizado como um diagnóstico de alerta precoce de comportamento semelhante a bolhas e tem demonstrado ser um método eficaz. [Hu e Oxley 2018, Monschang e Wilfling 2020, Phillips et al. 2015a]

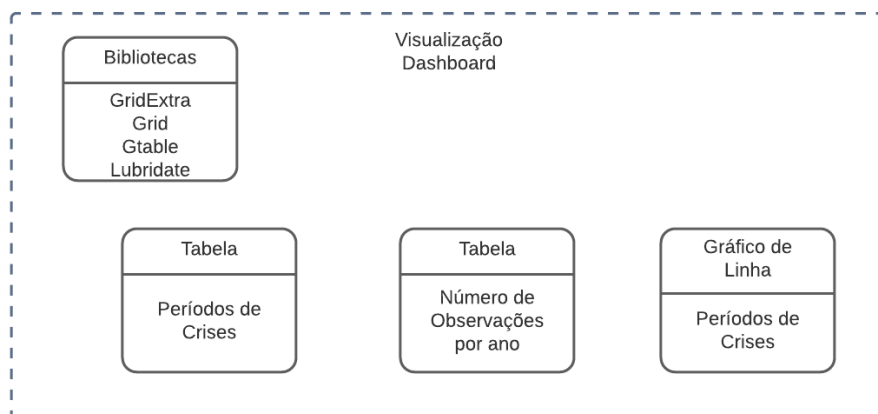


Figura 6: *Visualização*

5. Visualização - Na fase de visualização de dados, almejamos traduzir informações complexas ou abstratas de maneira gráfica e intuitiva, com o intuito de facilitar a compreensão, análise e interpretação. Através da perspectiva de *Dashboard* (painel de dados) no ambiente do *Databricks*, foi possível criar de modo tempestivo e objetivo um painel de dados que oferece suporte à tomada de decisões dos usuários, eliminando a necessidade de proficiência em programação. Os gráficos incorporam informações relativas aos períodos dos dados de entrada e aos períodos de bolhas financeiras resultantes da fase de *machine learning*, como descrito na Figura 6.

3.2 Modelo de *Machine Learning*

Utilizamos o método *General Supremum Augmented Dickey-Fuller* (GSADF), proposto em PSY (2020), para testar e identificar períodos de comportamento explosivo de preços, incluindo o início e fim do período. Uma vantagem desse método é sua independência em relação às informações financeiras e econômicas.

Este método utiliza evidências de comportamento explosivo de preços como *proxy* (substituto da real variável de interesse) para detecção de bolhas financeiras, o que pode levar a interpretações equivocadas. Por exemplo, se um determinado índice de mercado que está positivamente correlacionado com o preço do ativo está crescendo inesperadamente mais rápido do que antes, o método GSADF pode levar a conclusões errôneas de que há uma bolha de ativos [Escobari et al. 2017]. Ainda assim, o procedimento PSY hoje é o mais utilizado para detecção de bolhas, pois apresenta uma performance superior comparado a outros procedimentos, como por exemplo, sup-ADF-style, CUSUM (*cumulative sum control chart*), PWY, PSY^{sign-based} [Monschang e Wilfling 2020, Phillips et al. 2015a,b], pois o procedimento PSY mitiga o impacto potencial da heterocedasticidade.

A eficácia do procedimento PSY, aliada à sua disponibilidade em ferramentas econométricas, justifica a escolha deste método para a detecção de bolhas. O procedimento PSY estima um modelo de regressão específico para realizar essa detecção, proporcionando uma abordagem robusta na identificação de comportamentos anômalos nos preços dos ativos que podem indicar a presença de bolhas financeiras. Neste estudo, o seguinte modelo de regressão é estimado :

$$\Delta y_t = \hat{\alpha} + \hat{\beta} y_{t-1} + \sum_{i=1}^k \hat{\gamma}_i \Delta y_{t-i} + \hat{\varepsilon}_t, \quad (1), \text{ onde}$$

Δy_t : Representa a variação no valor da variável de interesse y no período de tempo t. Em termos mais simples, é a mudança nessa variável de um período para o próximo.

$\hat{\alpha}$: Refere-se a um coeficiente que constitui uma constante. Essencialmente, é um parâmetro que ajusta o modelo, garantindo um alinhamento adequado com os dados observados.

βy_{t-1} : Aqui, β é um coeficiente que multiplica o valor da variável y no período anterior (y_{t-1}). Isso sugere que o valor atual é influenciado pelo valor passado.

$\sum_{i=1}^k \gamma_i \Delta y_{t-i}$: Esta parcela envolve uma soma (Σ) que se estende por k períodos anteriores ($i=1$ até k). $\hat{\gamma}_i$ são coeficientes associados às mudanças nos valores da variável nos períodos anteriores (Δy_{t-i}). Isso indica que o valor atual depende das mudanças em períodos anteriores.

ε_t : Representa o erro no modelo, ou seja, a parte que não pode ser explicada pelas outras variáveis na equação. Este termo incorpora fatores não modelados ou não observados que contribuem para a variação não explicada.

Ao integrar esses componentes, a equação procura modelar como o valor corrente da variável (y_t) é influenciado pelo valor anterior (y_{t-1}) e pelas mudanças em períodos anteriores. Os coeficientes (α , β , $\hat{\gamma}_i$) são estimados a partir dos dados, ajustando o modelo aos padrões observados na série temporal. O procedimento PSY depende de estimativa repetida da regressão em subamostras dos dados de forma recursiva.

Quadro 2 - Variáveis de Pesquisa

Variável (y)	Fórmula	Fundamentação
Dividend yield	Dividendos pagos por ação / Cotação da ação x 100	[Caspi e Graham 2017, Escobari et al. 2017, Monschang e Wilfling 2020, Phillips et al. 2015a,b]
<i>Dividend yield</i> dos Juros sobre Capital Próprio(JSCP) ou <i>Dividend JSCP yield</i>	JSCP por ação / Cotação da ação x 100	[Caspi e Graham 2017, Escobari et al. 2017, Monschang e Wilfling 2020, Phillips et al. 2015a,b]

Para a estimação do modelo apresentado, utilizamos como dados de entrada as variáveis *Dividend Yield* e *Dividend JSCP yield*, como descrito em Caspi e Graham (2017), Escobari et al. (2017), Monschang e Wilfling 2020, Phillips et al. (2015a,b)

Dividend Yield é uma métrica financeira que expressa a relação entre os dividendos distribuídos por uma empresa e o preço de suas ações. Essa medida é

calculada como uma porcentagem e oferece aos investidores uma indicação do retorno em termos de dividendos que podem ser esperados em relação ao valor de mercado das ações. O *Dividend Yield* é utilizado como uma ferramenta de avaliação para comparar o retorno de dividendos entre diferentes investimentos. Uma taxa de *Dividend Yield* mais alta pode indicar um rendimento mais atraente para os investidores em busca de fluxo de caixa consistente.

Juros sobre capital próprio é outra forma de uma empresa distribuir o lucro entre os acionistas, titulares ou sócios. A justificativa para essa segregação reside no contexto empresarial brasileiro. Dividendos representam uma forma tradicional de distribuição de lucros de uma empresa aos seus acionistas. Entretanto, no Brasil, foi instituída uma prática alternativa denominada juros sobre capital próprio, que oferece vantagens fiscais ao possibilitar a redução do imposto de renda pago pelas empresas. Apesar de ambas as formas representarem distribuições de lucros, os dividendos são deduzidos diretamente do lucro, enquanto os juros sobre capital próprio influenciam na apuração dos lucros [Rodrigues da Silva e Kirch 2019].

O *Dividend yield* dos Juros sobre Capital Próprio (JSCP), ou simplesmente *Dividend JSCP yield*, é uma métrica financeira que expressa a relação entre os juros sobre capital próprio pagos por uma empresa e o preço de suas ações. Devido a essa particularidade contábil brasileira e à diferença substancial entre essas práticas, torna-se imperativo avaliar o comportamento de ambas.

4. Resultados

Este estudo propõe um *pipeline* de dados para detecção de bolhas financeiras no mercado brasileiro. Para superar a complexidade técnica envolvida, foi desenvolvido um pipeline de dados automatizado, que abrange desde a ingestão até a análise através do algoritmo PSY. Essa abordagem visa fornecer alertas significativos para bancos centrais, reguladores fiscais e agentes econômicos, simplificando o processo por meio de um painel visual intuitivo, facilitando a detecção de potenciais crises financeiras. A implementação foi realizada com a ferramenta Databricks.

Nesta seção as saídas do pipeline de dados são apresentadas Além disso, analisaremos a eficácia de utilização no mercado brasileiro.

4.1 Análise dos Dividend yield

**Contagem de Observações por Ano na B3:
Dividend Yield**

Ano	Contagem	Ano	Contagem
1 1996	59	15 2010	192
2 1997	70	16 2011	198
3 1998	46	17 2012	171
4 1999	36	18 2013	158
5 2000	33	19 2014	164
6 2001	41	20 2015	157
7 2002	43	21 2016	128
8 2003	34	22 2017	115
9 2004	47	23 2018	175
10 2005	68	24 2019	160
11 2006	88	25 2020	125
12 2007	146	26 2021	187
13 2008	221	27 2022	168
14 2009	169	28 2023	135

Figura 7: Contagem de observações por ano na B3 - Dividend Yield

O presente estudo utilizou dados de dividend yield das empresas listadas na B3, entre o período de 1996 e 2023, totalizando 3334 observações. O mercado brasileiro possui particularidades que o diferenciam dos mercados principais, como pequeno número de empresas negociando, baixa liquidez dos títulos, intervenção e participação governamental, e alta concentração de ações [De Amorim et al. 2021]. Tais características podem diminuir a eficiência do algoritmo de *machine learning*, especialmente nos primeiros 10 anos da amostra, conforme mostrado na Figura 7.

**Episódios de Crises das Empresas Listadas na B3:
Dividend Yield**

start	end
1 2008-03-01	2008-03-01
2 2008-09-01	2009-04-01
3 2010-05-01	2012-09-01

Figura 8: Episódios de Crises das empresas Listadas na B3 - Dividend Yield

A Figura 8 traça a relação preço/dividendo das empresas listadas na B3. As datas utilizadas se referem às datas de aprovação dos dividendos e não necessariamente da distribuição ou apuração, logo, essas datas podem representar períodos mais longos de comportamento financeiro.



Figura 9: Episódios de Crises das Empresas listadas na B3 - Dividend Yield

O procedimento PSY identificou dois períodos de bolhas. Como pode ser visto nas Figuras 8 e 9, ocorreu no início de 2008 um período de leve crescimento do dividend yield, seguido por um forte crescimento entre o final de 2008 e início de 2009. O contexto econômico era de abundância dos fluxos de capitais (ligado ao ciclo de liquidez direcionado aos países emergentes entre 2003 e 2007) e de resultados positivos nas transações comerciais e correntes do Brasil com o resto do mundo. Contudo, no final de 2009, a crise do subprime atingiu o país. No contexto brasileiro, as medidas anticíclicas, especialmente as relacionadas a políticas macroeconômicas expansionistas, foram cruciais para enfrentar a crise. Em 2009, observou-se uma recuperação na economia brasileira [Lima e Deus 2013].

O segundo período de crescimento do dividend yield aconteceu entre 2010 e 2012. Este período coincide com o início da crise das commodities (2010). Em um primeiro momento, a crise foi atenuada por políticas de estímulos fiscais do governo federal, por meio, principalmente, de desonerações tributárias, e de apoio aos bancos públicos, como fontes estratégicas para a sustentação do crédito para investimentos de longo prazo. Porém, a partir de meados de 2013 começou a se caracterizar uma reversão importante, quando o baixo dinamismo em termos de expansão do nível de atividade passou a ser acompanhado pela deterioração das métricas macroeconômicas [Arestis et al. 2017].

O modelo não detectou bolha financeira nos primeiros 12 anos da amostra, o que pode ser explicado pelo contexto macroeconômico brasileiro, que conviveu nos anos 90 com baixo crescimento econômico, alta inflação [Pinheiro et. al. 1999] e um aumento expressivo no déficit da conta corrente [Frizo e Lima, 2014], além de um mercado de capitais incipiente [De Amorim et al. 2021].

De acordo com os resultados apresentados utilizando *dividend yield*, demonstrou-se que o pipeline de dados é eficaz para automatizar o processo de ingestão, transformação, carga dos dados, *machine learning* e construção de gráficos que facilitam a identificação de bolhas financeiras das empresas listadas na B3.

4.2 Análise dos *Dividend JSCP yield*

Contagem de Observações por Ano na B3:
Dividend JSCP yield

Ano	Contagem	Ano	Contagem
1 1997	43	15 2011	136
2 1998	77	16 2012	142
3 1999	88	17 2013	169
4 2000	157	18 2014	166
5 2001	164	19 2015	164
6 2002	155	20 2016	160
7 2003	171	21 2017	190
8 2004	200	22 2018	206
9 2005	217	23 2019	246
10 2006	200	24 2020	176
11 2007	174	25 2021	265
12 2008	115	26 2022	260
13 2009	123	27 2023	124
14 2010	146		

Figura 10 : Contagem de observações por ano na B3 - *Dividend JSCP yield*

O presente estudo utilizou dados de *Dividend JSCP yield* das empresas listadas na B3, entre o período de 1997 e 2023, totalizando 4434 observações. Não foram encontradas distribuição de JSCP (Juros sobre capital próprio) no ano de 1996. Assim como observamos nos resultados para *Dividend Yield*, as características do mercado brasileiro também podem influenciar na eficácia dos modelos de machine learning, especialmente nos 10 primeiros anos da amostra.

Episódios de Crises das Empresas Listadas na B3:
Dividend JSCP yield

	start	end
1	2008-03-01	2008-03-01
2	2008-09-01	2009-04-01
3	2010-05-01	2012-09-01

Figura 11 : Episódios de Crises das empresas Listadas na B3 - *Dividend JSCP yield*

A análise do *Dividend JSCP yield* apresentou uma tendência semelhante ao do *Dividend Yield* que é explicado pelo fato de ambas variáveis serem funções diretamente proporcionais ao lucro das empresas. Os resultados demonstram dois períodos de crescimento, sendo o primeiro associado à abundância de fluxos de capitais entre 2003 e 2007, seguido pela crise do subprime em 2009, que foi enfrentada por meio de medidas anticíclicas. O segundo período coincide com a crise das commodities em 2010, sendo mitigado inicialmente por estímulos fiscais, mas seguido por uma reversão a partir de meados de 2013 [Arestis et al. 2017].



Figura 12: Episódios de Crises das Empresas listadas na B3 - *Dividend JSCP Yield*

Durante os primeiros 12 anos da análise, o algoritmo de *machine learning* não conseguiu detectar bolhas financeiras de forma semelhante ao que foi observado ao analisar o *Dividend Yield*. Essa limitação pode ser atribuída ao contexto macroeconômico brasileiro nas décadas de 90 e 2000. Nesse período, o país enfrentou desafios como baixo crescimento econômico, alta inflação e um mercado de capitais em estágio inicial de desenvolvimento [Pinheiro et al. 1999; De Amorim et al. 2021]. Esses fatores combinados podem ter contribuído para a dificuldade do algoritmo em identificar bolhas financeiras durante esse período específico.

As análises dos dados de *Dividend JSCP yield* corroboram os resultados apresentados utilizando *Dividend yield*, o que ressalta a eficácia do pipeline de dados para a identificação de bolhas financeiras nas empresas listadas na B3.

4.3 Análise dos dados combinados (*Dividend JSCP yield + Dividend yield*)

Contagem de Observações por Ano na B3:

Total = Dividendos + JSCP

Ano	Contagem	Ano	Contagem
1 1996	59	15 2010	338
2 1997	113	16 2011	334
3 1998	123	17 2012	313
4 1999	124	18 2013	327
5 2000	190	19 2014	330
6 2001	205	20 2015	321
7 2002	198	21 2016	288
8 2003	205	22 2017	305
9 2004	247	23 2018	381
10 2005	285	24 2019	406
11 2006	288	25 2020	301
12 2007	320	26 2021	452
13 2008	336	27 2022	428
14 2009	292	28 2023	259

Figura 13: Contagem de observações por ano na B3 - Dados combinados

Com caráter comprobatório, o presente estudo empregou dados de rendimento de *Dividend yield* e *Dividend JSCP yield* (Juros Sobre Capital Próprio) de forma combinada. A amostra abrange as empresas listadas na B3, no período entre 1996 a 2023, totalizando 7767 observações.

Episódios de Crises das Empresas Listadas na B3:

Total = Dividend + JSCP

	start	end
1	2008-03-01	2008-03-01
2	2008-09-01	2009-04-01
3	2010-05-01	2012-09-01

Figura 14 : Episódios de Crises das empresas Listadas na B3 - Dados Combinados

Assim como nos resultados seccionados, a análise conjunta do *Dividend Yield* e do *Dividend JSCP yield* revela dois períodos de crescimento. Este resultado era esperado porque ambas as variáveis são funções do lucro líquido das empresas.

O primeiro período de crescimento generalizado está associado à abundância de fluxos de capitais entre 2003 e 2007, seguido pela crise do subprime em 2009, que foi

enfrentada por meio de medidas anticíclicas. O segundo período coincide com a crise das commodities em 2010, sendo mitigado inicialmente por estímulos fiscais, mas seguido por uma reversão a partir de meados de 2013 [Arestis et al. 2017]. Como nos casos anteriores, o algoritmo de *machine learning* não identificou bolhas financeiras nos primeiros 12 anos da amostra.



Figura 15: Episódios de Crises das Empresas listadas na B3 - Dados combinados

As análises dos dados combinados corroboram os resultados apresentados utilizando *Dividend yield* e *Dividend JSCP yield* o que ressalta a eficácia do pipeline de dados para a identificação de bolhas financeiras nas empresas listadas na B3.

4.4 Painel de detecção de bolhas e episódios de crises

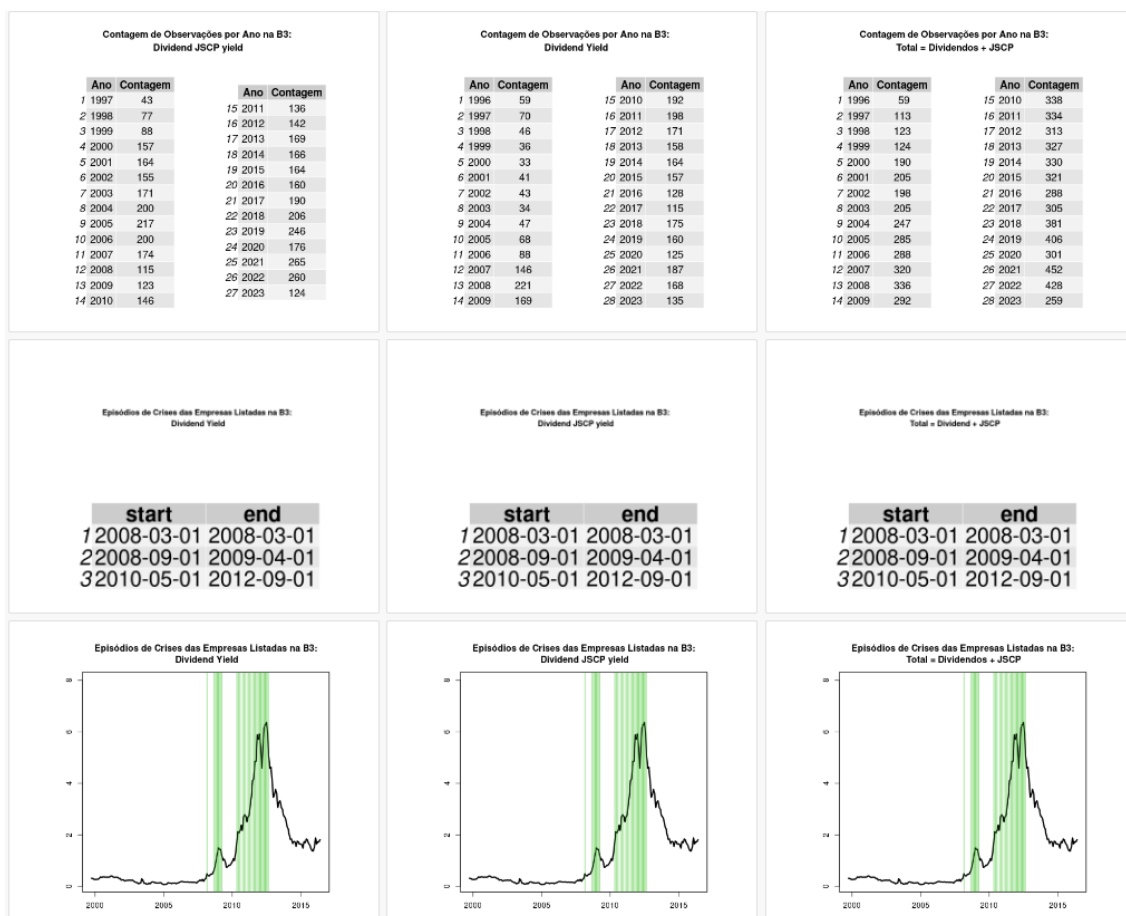


Figura 16: Painel de bolhas e crises de empresas listadas na B3

Os gráficos gerados neste estudo integram um *dashboard* com gráficos e métricas que facilitam a identificação de bolhas financeiras das empresas listadas na B3, conforme a figura 16. O painel inclui os dados referentes às análises de *Dividend Yield*, *Dividend JSCP Yield* e os dados combinados das duas variáveis. O painel foi construído na ferramenta databricks o que permite automatizar a execução do pipeline de dados atualizando o painel com dados atualizados. Com isto, o usuário final não necessita de conhecimentos específicos nas áreas de programação e estatística, possibilitando-os ter um foco na análise dos resultados.

5. Considerações Finais

O presente trabalho propôs um pipeline de dados automatizado para a detecção de bolhas financeiras no mercado de ativos brasileiro, utilizando o algoritmo de *machine learning* conhecido como procedimento PSY. O objetivo principal era fornecer uma ferramenta eficaz para bancos centrais, reguladores fiscais e agentes econômicos identificarem precocemente sinais de comportamento especulativo e potenciais crises financeiras, visando a estabilidade econômica.

O pipeline desenvolvido mostrou-se eficaz na automação do processo de ingestão, transformação e carga de dados, além da aplicação do algoritmo PSY. A criação de um painel visual intuitivo permitiu uma análise fácil e rápida das empresas listadas na B3, facilitando a identificação de períodos de bolhas financeiras. A utilização da ferramenta Databricks proporcionou uma implementação eficiente e escalável do pipeline, demonstrando sua viabilidade prática.

A análise dos resultados revelou dois períodos de crescimento do *Dividend Yield* e do *Dividend JSCP Yield*, coincidindo com eventos macroeconômicos significativos, como a crise do subprime em 2009 e a crise das commodities em 2010. A eficácia do pipeline foi evidenciada, uma vez que conseguiu identificar esses períodos de bolhas financeiras, fornecendo informações valiosas para a compreensão dos movimentos do mercado.

Apesar dos resultados positivos, é importante destacar algumas limitações e desafios enfrentados durante a implementação do pipeline. O contexto macroeconômico brasileiro nas décadas de 90 e 2000, caracterizado por baixo crescimento econômico e alta inflação, pode ter influenciado na ausência de detecção de bolhas nos primeiros 12 anos da amostra. Além disso, a baixa liquidez e a concentração de ações no mercado brasileiro apresentam desafios adicionais para a eficácia do algoritmo de *machine learning*.

Para aprimorar a detecção de bolhas financeiras no mercado brasileiro, é sugerido testar o modelo com diferentes variáveis financeiras e explorar outros algoritmos de *machine learning*. A inclusão de indicadores econômicos abrangentes e a adaptação do modelo a diversos contextos de mercado fortalecerão sua capacidade de identificar bolhas financeiras. Além disso, a incorporação proativa de inovações tecnológicas e a consideração de novas fontes de dados são fundamentais para garantir a eficácia e o aprimoramento contínuo do sistema ao longo do tempo.

Em conclusão, o pipeline de dados proposto neste trabalho apresenta uma contribuição significativa para a detecção de bolhas financeiras no mercado de ativos brasileiro. A automação do processo, aliada à utilização da técnica PSY, fornece uma ferramenta valiosa para os agentes do mercado financeiro e reguladores, permitindo uma resposta mais eficiente a potenciais crises. Embora desafios e limitações persistam, a evolução contínua e adaptação do pipeline podem fortalecer sua utilidade no monitoramento e prevenção de bolhas financeiras, contribuindo assim para a estabilidade econômica do país.

Referências

- Arestis, P., Baltar, C., e Prates, D. (Eds.). (2017). *The Brazilian Economy since the Great Financial Crisis of 2007/2008*. London: Palgrave Macmillan.
- Bourgard, B. e Gomes, C. (2017). As variáveis econômicas no Brasil e o PIB: uma análise em períodos de crises financeiras através da correlação de Pearson. *Almanaque Multidisciplinar de Pesquisa*, 4(2).

- Bragagnolo, G. (2020). Pecúária bovina no Brasil e disfuncionalidades do mercado financeiro: um estudo sobre os impactos no valor de mercado dos frigoríficos brasileiros de capital aberto decorrente do aumento da demanda chinesa em virtude da peste suína africana (Doctoral dissertation).
- Caspi, I. e Graham, M. (2018). Testing for bubbles in stock markets with irregular dividend distribution. *Finance Research Letters*, 26, 89-94.
- Chaim, P. e Laurini, M. (2019). Is Bitcoin a bubble?. *Physica A: Statistical Mechanics and its Applications*, 517, 222-232.
- De Amorim, G., Lima, N. e Júnior, A. (2021). Distribuição de Dividendos e Valor de Empresas Listadas na B3. *Advances in Scientific and Applied Accounting*, p. 003-018.
- Dias Junior, E. (2022). Crise financeira e sanitária da Covid-19: análise de impacto financeiro nas indústrias brasileiras (Doctoral dissertation).
- Deloitte. As maiores crises em ordem cronológica. 2020. <https://www2.deloitte.com/content/dam/Deloitte/br/Documents/risk/Deloitte-infografico-riscos-RA-2020.pdf>, Acessado em: 04 de set. de 2023
- Escobari, D., Garcia, S. e Mellado, C. (2017). Identifying bubbles in Latin American equity markets: Phillips-Perron-based tests and linkages. *Emerging Markets Review*, 33, 90-101.
- Espindola, R. (2015). A crise financeira e a política monetária no Brasil (Doctoral dissertation).
- Frizo, P. e Lima, R. (2014). Efeitos da flutuação dos preços das commodities no fluxo de investimento estrangeiro direto no Brasil. *Revista de Economia Contemporânea*, 18, 393-408.
- Hu, Y. (2023). A review of Phillips-type right-tailed unit root bubble detection tests. *Journal of Economic Surveys*, 37(1), 141-158.
- Hu, Y. e Oxley, L. (2018). Bubble contagion: Evidence from Japan's asset price bubble of the 1980-90s. *Journal of the Japanese and International Economies*, 50, 89-95.
- Kufel, J., Bargiel-Laczek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., e Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics*, 13(15), 2582.
- Lima, T. e Deus, L. (2013). A crise de 2008 e seus efeitos na economia brasileira. *Revista Cadernos de Economia*, 17(32), 52-65.
- Monschang, V. e Wilfling, B. (2021). Sup-ADF-style bubble-detection methods under test. *Empirical Economics*, 61, 145-172.
- Phillips, P., Shi, S. e Yu, J. (2015a). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International economic review*, 56(4), 1043-1078.

- Phillips, P., Shi, S. e Yu, J. (2015b). Testing for multiple bubbles: Limit theory of real-time detectors. *International Economic Review*, 56(4), 1079-1134.
- Phillips, P., e Shi, S. (2018). Financial bubble implosion and reverse regression. *Econometric Theory*, 34(4), 705-753.
- Phillips, P. e Shi, S. (2020). Real time monitoring of asset markets: Bubbles and crises. In *Handbook of statistics* (Vol. 42, pp. 61-80). Elsevier.
- Pinheiro, A., Giambiagi, F. e Gostkorzewicz, J. (1999). O desempenho macroeconômico do Brasil nos anos 90.
- Rodrigues da Silva, A. e Kirch, G. (2019). Efeito clientela no setor elétrico brasileiro e suas possibilidades de arbitragem. *Revista de Administração da UEG*, 10(3).
- Shi, S. e Phillips, P. (2022). *Econometric Analysis of Asset Price Bubbles* (No. 2331). Cowles Foundation for Research in Economics, Yale University.
- Usmani, S. e Shamsi, J. (2021). News sensitive stock market prediction: literature review and suggestions. *PeerJ Computer Science*, 7, e490.