Uma Abordagem Bibliométrica e de Machine Learning para Estudos DSGE

Diego Batalha* Angelo Alves†

2021, v-1.0

Resumo

O presente artigo tem a intenção de predizer qual a melhor revista para ser publicado um artigo produzido que utilizou DSGE para modelos macroeconômicos, dado o seu abstract. A hipótese é que cada revista possui modelos preferênciais. O trabalho se mostra relevante pois o tempo de apreciação e avaliação de artigos pelas comissões e pelo avaliador é grande e qualquer forma de diminuir possíveis reprovações ou rejeições do artigo é um ganho. Neste trabalho, fez-se uma revisão da literatura: dos artigos mais recentes que fizeram uso de modelos de equilíbrio geral estocástico e dinâmico; dos artigos que explicam a análise bibliométrica para observar as principais tendências e mapeamento científico; e dos artigos que explicam o problema de classificação aplicado ao Processamento de Linguagem Natural (NLP). Para isso, foi utilizado técnicas de *Machine Learning* como *Decision Tree* e *Random Forest*.

Palavras-chaves: DSGE, bibliometria, *Machine Learning*, *random forest*, *decision tree*, macroeconomia.

Introdução

O *Dynamic Stochastic General Equilibrium* (DSGE) é um dos modelos econômicos mais utilizados em macroeconomia, principalmente por sua fundamentação microeconômica. Desta forma, considera que as relações entre as variáveis macroeconômicas são frutos de decisões ótimas de agentes econômicos, tais como famílias, firmas e autoridades monetária e fiscal, operando sob as restrições impostas pelo ambiente na qual cada agente opera. O modelo DSGE tem a capacidade de incorporar características empíricas da economia real como por exemplo, a adoção da hipótese de poder de mercado das firmas. Outra característica comumente empregada é a admissão de rigidez de preços e salários, que influenciam no efeito da política monetária sobre a economia. As fricções reais - tais como custos de ajustamento do capital, utilização variável da capacidade instalada e formação de hábito no consumo, como acelerador financeiro, enriquecem o modelo e o tornam mais próximo dos fenômenos econômicos.

*silva.diego_6@posgraduacao.uerj.br
†alves.angelo@posgraduacao.uerj.br

Em síntese, esses modelos tem como objetivos identificar as flutuações das variáveis macroeconômicas, entender os meios de propagação dos choques, prever os impactos das mudanças de políticas econômicas e até mesmo realizar previsões no futuro de variáveis chaves da economia. Para montar um modelo DSGE de forma eficiente, a estimação e inferência devem levar em conta a necessidade de conjugar coerência teórica e implementação empírica.

Para analisar os modelos que utilizaram DSGE, esse artigo fez uso de metodologias bibliométricas, que são consideradas úteis como ferramentas de apoio[1] ao acompanhamento da evolução da ciência. Parte dessa disseminação se deve à abundância de dados e à facilidade de acessibilidade a um grande número de ferramentas de processamento bibliometrico. (Zuccala, 2016; Rousseau e Rousseau, 2017). Este artigo também se propõe a classificá-los, dado o seu abstract, buscando o melhor veículo para submetê-lo e diminuindo as chances de rejeição. Portanto, conforme o supracitado, temos um problema de classificação

O artigo foi estruturado da seguinte forma: (i) a primeira parte, foi feito uma breve introdução às definições de DSGE, bibliometria e modelos de classificação; (ii) em seguida, detalhou-se a metodologia e uma estatística descritiva dos dados; e, (iii) por último, os resultados foram apresentados e as prospecções foram feitas para futuros estudos.

Os modelos DSGE

Sobre a modelagem estocástica, que é a abordagem teórica mais comum na literatura atual, Eichenbaum *et al.* (2020) utilizam um modelo de ciclo de negócios real (RBC) ampliado com um modelo SIR epidemiológico, para examinar a dinâmica das epidemias. Vários estudos estendem os modelos RBC e NK típicos com rigidez e atritos adicionais, para permitir dinâmicas econômicas mais complexas. Um aspecto que tem sido menos enfatizado é a heterogeneidade entre as famílias, que tem implicações importantes para a análise do bem-estar, respostas ao choque e política governamental (Galí *et al.*, 2007). Zhang *et al.* (2021) incluem heterogeneidade ao assumir famílias ricardianas e não ricardianas; no entanto, seu modelo é calibrado para se ajustar à economia da China. A literatura atual também se concentra geralmente nos EUA e na área do euro.

De forma a trazer os estudos mais recentes sobre DSGE, em julho de 2020, além das previsões econômicas globais produzidas pelas instituições financeiras internacionais, apenas alguns estudos estavam tentando modelar as consequências macroeconômicas globais do COVID-19. Os estudos da Organização Mundial do Comércio (2020), Maliszewska *et al.* (2020) e o World Bank (2020) utilizam modelos de Equilíbrio Geral Computável (CGE) e focam principalmente no impacto da mortalidade, morbidade e aumento dos custos de produção nas economias. Um estudo do FMI (2020), que utiliza um modelo semi-estrutural Dynamic Stochastic General Equilibrium (DSGE), também inclui perturbações nos mercados financeiros.

McKibbin e Fernando (2020a) desenvolveram cenários de pandemia global para formuladores de políticas em uma série de países em fevereiro de 2020. Eles usaram a experiência histórica de outras epidemias globais importantes para explorar sete cenários diferentes para a economia mundial. Estimaram a transmissão epidemiológica entre os países com base em vários indicadores e, em seguida, usaram esses resultados epidemiológicos para projetar um conjunto de choques econômicos. Esses choques foram então aplicados

[1] além de ser essencial para a tomada de decisão no estabelecimento de prioridades de pesquisa e tecnologia e recompensa da excelência científica.

ao modelo econômico global do G-Cubed. A análise forneceu uma série de estimativas das prováveis consequências macroeconômicas do COVID-19 sem intervenções de saúde pública. A pesquisa foi atualizada por McKibbin e Fernando (2020c) em junho de 2020. O segundo grande artigo usou dados reais para a pandemia COVID-19 e, em seguida, aplicou-os juntamente com suposições sobre as diferentes durações das ondas pandêmicas e políticas econômicas e de saúde já anunciadas pelos governos.

McKibbin e Fernando (2021) estende a análise para uma nova versão do modelo G-Cubed, com foco nas economias asiáticas em uma estrutura global. Também avalia opções de políticas plausíveis para apoiar a recuperação econômica. Primeiro atualiza as estimativas do impacto macroeconômico global da pandemia com dados fornecidos até novembro de 2020, antes de avaliar como as opções de políticas potenciais poderiam reduzir as consequências macroeconômicas adversas da pandemia.

McKibbin e Fernando (2021) estendeu a abordagem de McKibbin e Fernando (2020a, c) para explorar o impacto da pandemia COVID-19 nas economias asiáticas e quatro diferentes respostas políticas foram encontradas: (i) um aumento nos pagamentos de transferência para as famílias; (ii) gastos adicionais do governo em bens e serviços; (iii) aumento nos gastos com infraestrutura; e, (iv) uma resposta de saúde pública muito melhor, incluindo a implantação rápida de uma vacina. Esses resultados pretendem ser ilustrativos, uma vez que as magnitudes exatas de qualquer política em uma economia específica dependerão dos detalhes precisos do pacote.

Os resultados de McKibbin e Fernando (2021) sugerem que a maioria dos benefícios viria de uma resposta robusta de saúde pública e rápida implantação de uma vacina. [1]

A Pesquisa Bibliométrica

Bibliometria refere-se à "aplicação de métodos matemáticos e estatísticos a livros e outras formas de comunicação escrita" (Pritchard, 1969). Alguns artigos que revisam as definições bibliométricas são Chellappandi e Vijayakumar (2018) e Hlavcheva *et al.* (2019). Visões gerais abrangentes do campo podem ser encontradas em Glänzel (2003) e Rousseau *et al.* (2018).

Outra abordagem para capturar tendências na pesquisa bibliométrica é estudar os artigos publicados pelos principais periódicos da área. Nesse sentido, Schoepflin e Glanzel coletaram artigos publicados em 1980, 1989 e 1997 na revista Scientometrics e classificaram manualmente os registros recuperados em seis categorias[2](Schoepflin e Glänzel, 2001). Estudos de caso com essa abordagem se tornaram dominantes no período mais recente.

A outra abordagem prevalente para aquisição de dados nesses meta-estudos é o uso de palavras-chave para realizar pesquisas temáticas em bancos de dados bibliográficos. Ao revisar as tendências na pesquisa de métricas de informação, Bar-Ilan (2008) desenvolve uma consulta abrangente listando uma variedade de termos relacionados a métodos bibliométricos. Os dados foram extraídos da *Web of Science* (WoS), *Scopus*, *Google Scholar* e outros bancos de dados para os anos de 2000-2006, resultando em 598 artigos após a

[1] Os outros programas de políticas ajudam a aliviar o impacto macroeconômico da pandemia COVID-19 e, talvez, uma combinação de cada política teria um efeito significativo.

[2] 1. Teoria bibliométrica, modelos matemáticos e formalização de leis bibliométricas, 2. Estudos de caso e artigos empíricos, 3. Artigos metodológicos incluindo aplicações, 4. Engenharia de indicadores e apresentação de dados, 5. Abordagem sociológica da bibliometria, sociologia da ciência, 6. Política científica, gestão científica e discussões gerais ou técnicas

filtragem. Verificou-se que tópicos tradicionais como análise de citações, fator de impacto e pesquisa de índice h continuam em ascensão, mas também tópicos mais recentes como webometria, mapeamento e visualização e acesso aberto estão sendo introduzidos como tópicos recorrentes na bibliometria. (Mooghali *et al.*, 2011; Ellegaard e Wallin, 2015).

Com um conjunto de dados de 23.296 artigos obtidos de uma lista mais longa de termos de pesquisa, Maltseva e Batagelj (2020) aplicam a análise de citações para descobrir caminhos evolutivos, ou cadeias de citações, entre pesquisadores de bibliometria. Eles estão particularmente interessados em descobrir padrões de colaboração, descobrindo que o número de artigos publicados sobre bibliometria dobra a cada 8 anos e que os artigos colaborativos com três ou mais autores estão aumentando em comparação com a tendência decrescente de artigos de autoria única.

Seguindo uma abordagem baseada em palavras-chave para extração de dados, mas com o objetivo de encontrar diferenças entre a pesquisa bibliométrica dentro e fora da informação e da biblioteconomia, foi encontrado o trabalho de Jonkers e Derrick (2012) que estudou 3.852 artigos bibliométricos publicados entre 1991 e 2010 e comparou citações e autor de artigos publicados em periódicos de diversos assuntos. Mais notavelmente, González-Alcaide (2021) faz a comparação considerando as redes de colaboração do autor e diferentes níveis de domínio, incluindo ciências sociais, ciências biológicas e medicina, tecnologia, ciências físicas, ciências multidisciplinares e artes e humanidades como categorias de avaliação, observando poucos laços de colaboração com o núcleo da pesquisa bibliométrica e equipes dispersas trabalhando de forma independente.

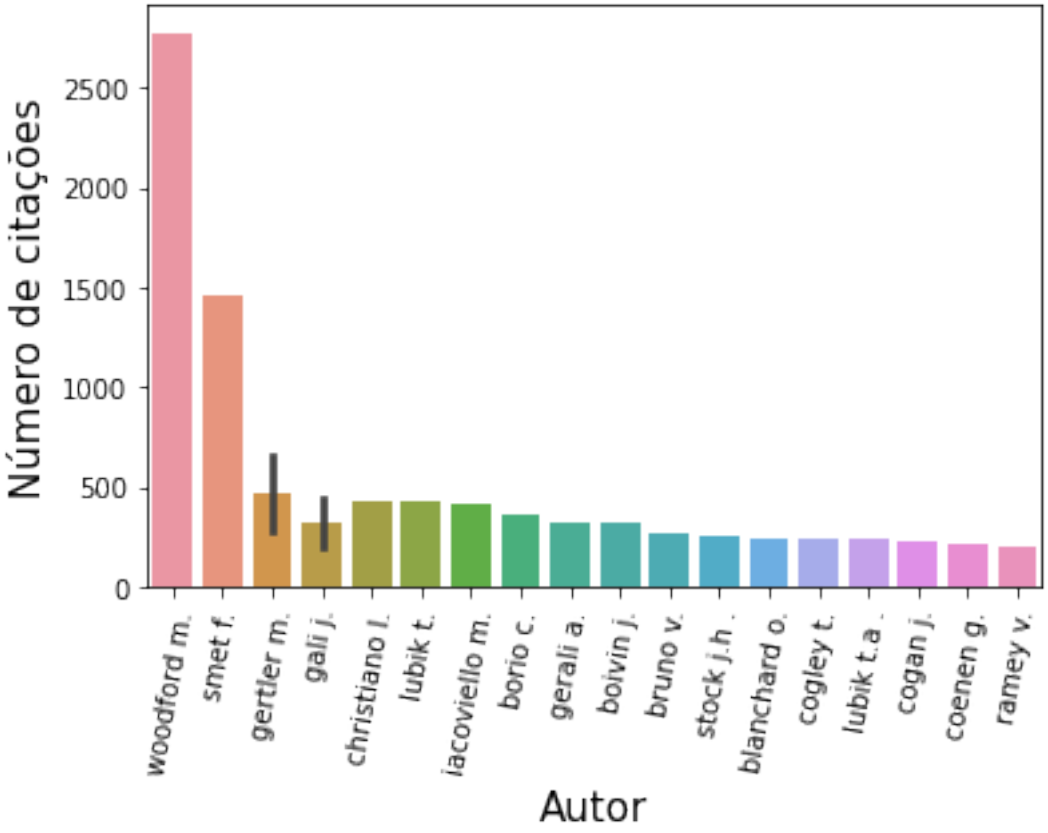
Metodologia da Pesquisa e Análise Descritiva

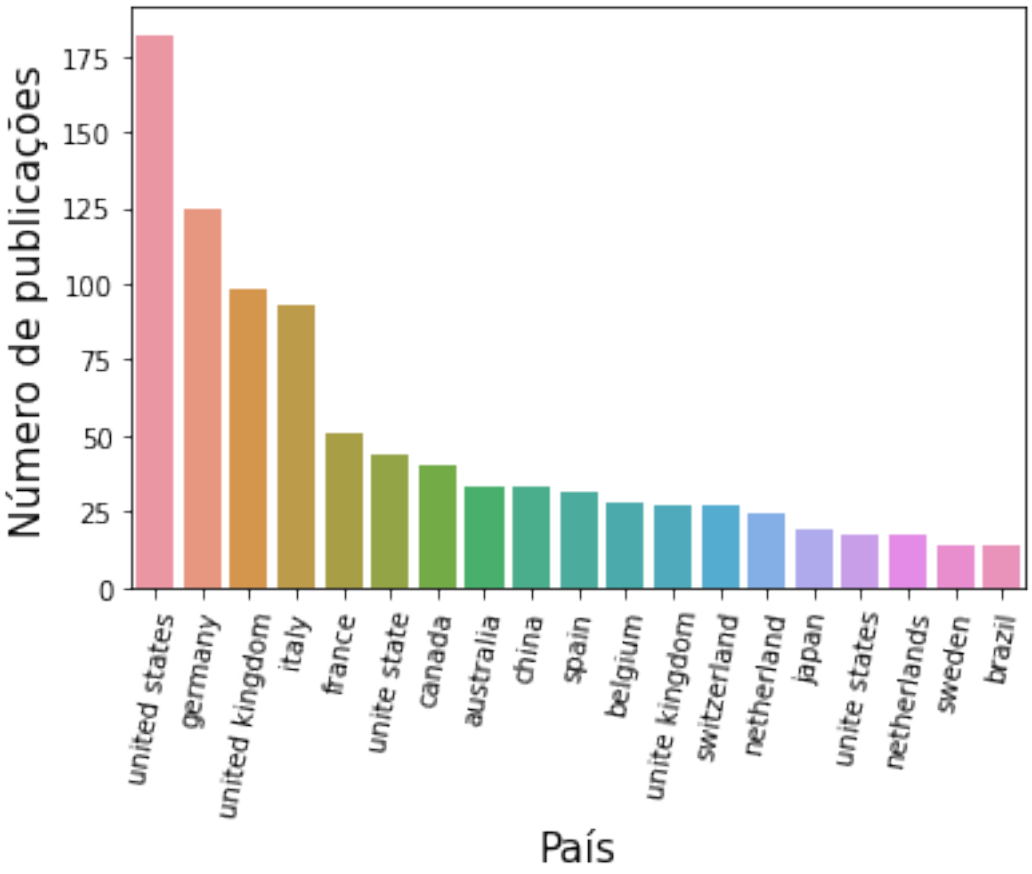
Análise Exploratória, Pacotes Python e Pré-processamento do Banco de Dados

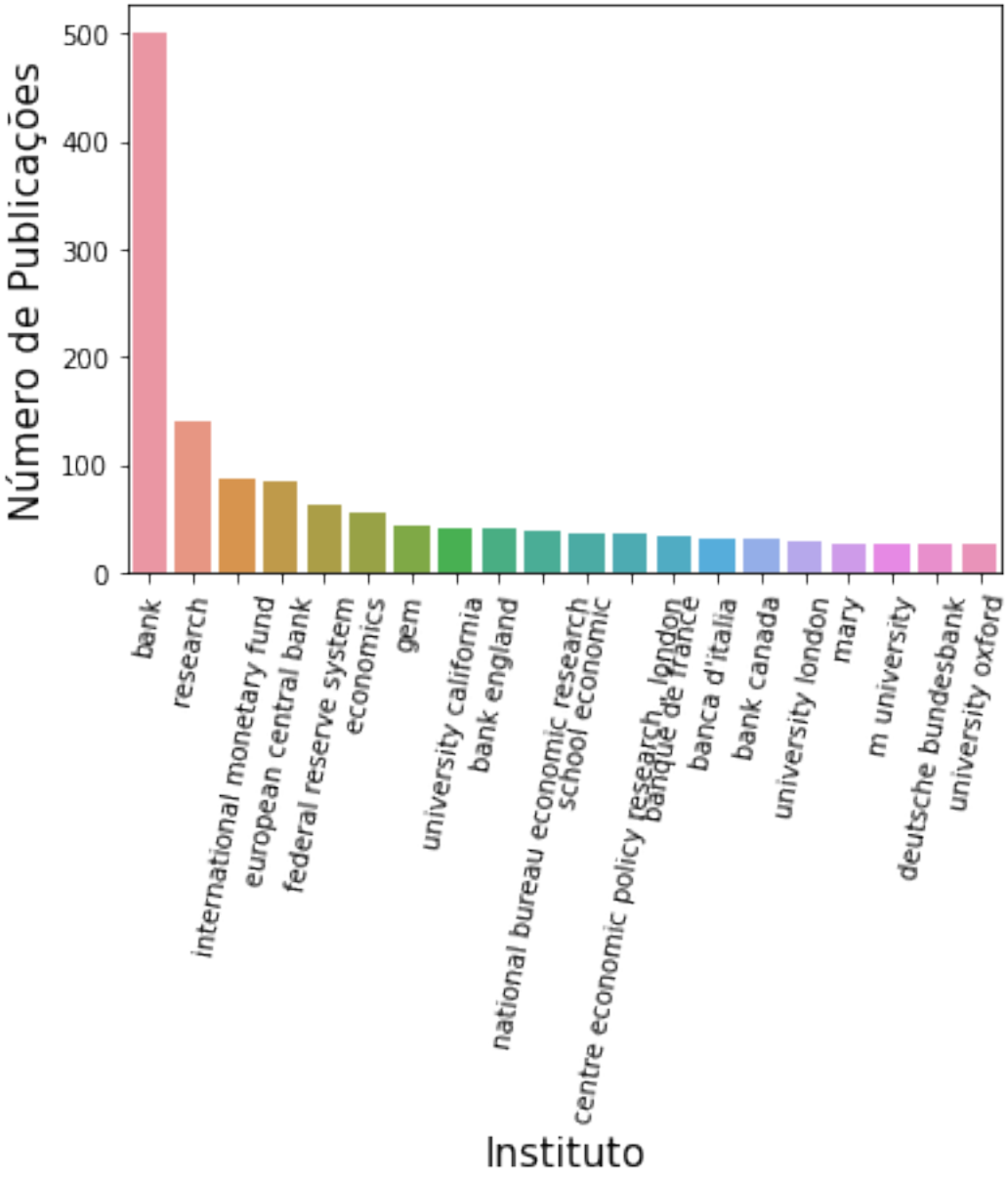
A intenção dos modelos utilizados nesse trabalho são predizer qual a melhor revista para ser publicado um artigo produzido, dado o seu abstract. A hipótese é que cada revista possui modelos preferênciais. O trabalho se mostra relevante pois o tempo de apreciação e avaliação de artigos pelas comissões e pelo avaliador é grande e qualquer forma de diminuir possíveis reprovações ou rejeições do artigo é um ganho. Portanto, há um prolema de classificação.

As bibliotecas *Numpy*, *Pandas* e *Spacy* foram utilizadas a fim de manusear os dataframes. Foi feito um pré-processamento padrão de *Natural Language Processing (NLP)* para retirar dos abstracts brutos os caracteres que não acrescentam maior interpretabilidade aos dados, desde espaços duplos até tokenizar todos as palavras e deixá-las em *lower case*.

O banco de dados foi baixado diretamente pela API do Scopus através do pacote *pybliometrics* desenolvido para *Python* (Rose and Kitchin 2019). Seguiu-se a abordagem baseada em palavras-chave - explicada na seção anterior, para extração de dados. A coleção principal inclui artigos de diferentes disciplinas, além da macroeconomia, e tipos de documentos. Para obter documentos que empregam explicitamente os termos pertinentes, realizou-se uma busca tópica com a consulta *s = ScopusSearch("DSGE","macroeconomics")*. Uma pesquisa por tópico recupera os registros correspondentes à consulta no título, resumo ou palavras-chave. Nenhuma restrição de tempo foi colocada nesse primeiro momento de busca de registros. Os dados foram recuperados em 29 de outubro de 2021, obtendo-se 7.765 registros, usando a função *ScopusSearch()*. Para a análise exploratório, será contabilizada as publicações por países, por autores e por institutos:


Número de citações
2500
2000
1500
1000
500
0
woodford m.
smet f.
gertler m.
gali j.
christiano l.
lubik t.
iacoviello m.
borio c.
gerali a.
boivin j.
bruno v.
stock j.h.
blanchard o.
cogley t.
lubik t.a.
cogan j.
coenen g.
ramey v.
Autor


Número de publicações
175
150
125
100
75
50
25
0
united states
germany
united kingdom
italy
france
unite state
canada
australia
china
spain
belgium
unite kingdom
switzerland
netherland
japan
unite states
netherlands
sweden
brazil
País



Realizou-se o pré-processamento dos abstracts dos 7.765 artigos do banco de dados que são: limpeza, integração, transformação, redução e discretização dos dados. (Bird *et. al*, 2009). No segundo momento, restringiu-se um pouco mais a base de dados, seja referente ao número de revistas, que foi restringida em 10 dentre as aproximadamente 400 revistas que mais publicam sobre DSGE, seja o tempo, que foi limitado dos anos 2000 até os dias atuais. Sendo assim, com um pouco mais de refino, chegou-se a base de dados final constituida de um total de 762 arquivos.

Do conjunto dos 762 artigos do banco de dados, 95% é composto de artigos de periódicos, 1,7% de revisões, 1,4% de papers de conferências e os demais de outros tipos, incluindo material editorial e capítulos de livros. Todos os tipos de registros foram incluídos nesta pesquisa, e serão referidos como "artigos" de forma genérica no restante deste trabalho. A lista completa de artigos, incluindo o ID do artigo do banco de dados, identificador do objeto do documento e uma etiqueta indicando se eles correspondem a qualquer um dos termos consultados, é oferecida como material complementar.

Definindo as Variáveis para o Modelo

Foi definido como variável explicativa a variavel categórica da revista em que determinado artigo foi publicado. Portanto teremos Veiculo como a Variavel explicada. Em seguida, definiu-se as variáveis explicativas como as palavras do abstract do artigo, através do método *Term Frequency–Inverse Document Frequency* (TF-IDF).

Primeiro, a Frequência do Termo (TF) é usado para medir quantas vezes um termo está presente em um documento. Mas, suponha que para um documento contendo 5000 palavras e a palavra α está presente no documento exatamente 10 vezes. É sabido que o comprimento total dos documentos pode variar muito, portanto, é possível que qualquer termo ocorra com mais frequência em documentos grandes em comparação com documentos pequenos. Assim, para corrigir esse problema, a ocorrência de algum termo em um documento é dividida pelo total de termos presentes naquele documento, para encontrar a frequência do termo. Na Frequência Inversa do Documento (IDF), quando o TF de um documento é calculado, pode-se observar que o algoritmo trata todas as palavras-chave igualmente, não importa se é uma *Stopword* como "de", o que não é correto. Todas as palavras-chave têm importância diferente. Assumindo que a *Stopword* "de"esteja presente em um documento 2.000 vezes, mas não tem utilidade ou tem um significado muito menor. A IDF atribui menor peso às palavras frequentes e atribui maior peso às palavras pouco frequentes. (Qaiser e Aly, 2018)

A próxima etapa envolve o uso de uma função de ajuste para resolver o problema de classificação, separando as variáveis explicadas e explicativas em treino e teste para validar os resultados adquiridos. Para dividir a base de dados, usou-se a função *train_test_split* do pacote *Scikit-Learn*. Para isso, utilizou-se os modelos de machine learning como *Decision Tree* e *Random Forest* e foi utilizado o *DummyClassifier* para balizá-los (*Scikit-Learn*). Utilizou-se, também, uma SEED igual 301 como *Random State* para tornar a comparação de resultados possível.

Metodologia dos Modelos de Machine Learning

O classificador de árvore de decisão (James *et.al.*, 2013) é um método que fornece as classes previstas como um caminho com resultados possíveis levando a cada classe. Cada ramo da árvore é uma variável particular.

Florestas aleatórias são algoritmos de aprendizado de máquina que usam um conjunto de árvores de decisão (chamadas de floresta) para prever um resultado (Breiman, 2001). Em resumo, o algoritmo cria árvores de decisão individuais (i) selecionando aleatoriamente um subconjunto do conjunto de treinamento; (ii) selecionando aleatoriamente um subconjunto de recursos em cada divisão; e, (iii) mantendo o recurso que diminui a incerteza da decisão. O algoritmo então repete esse processo várias vezes para criar uma floresta com árvores diferentes. Para problemas de classificação, as previsões de todas as árvores independentes são agregadas e a previsão mais popular é selecionada como o rótulo de classe prevista.

Normalmente, os dois métodos de classificação usam validação cruzada de dez vezes para dividir o conjunto de dados em treinamento e conjunto de teste e fazer previsões (James *et.al.*, 2013).

Dado os hiperparâmetros para o algoritmo. A variável random_state que nesse trabalho é igual a 301, é opcional na prática, mas garante a produção de uma saída consistente. Como há um componente aleatório no algoritmo, definir o *random_state* irá

certificar obter os mesmos resultados que os apresentados neste tutorial sempre que executar o código em Python.

Definir o *class_weight* como balanceado garante que ambos os valores da classe rótulo carregam o mesmo peso, o que é necessário. Conseqüentemente, equilibrar os pesos dos valores do rótulo da classe evita que o modelo superclassifique as previsões no valor do rótulo da classe com o maior número de amostras.

Resultado dos Modelos de Machine Learning

No modelo dummy, a acurácia com cross validation com $k = 5$ ficou no intervalo de $16,21\%$ e $18,43\%$ com valor médio em 17.09%. $<< Accuracycomdummystratified, 10 = [16.21\%, 18.43\%] >>$

Utilizou-se o *Randomized Grid Cross Validation* para o modelo de *Machine Learning Decision Tree* com um conjunto de valores para profundidade máxima como $[10, 15, 20, 30, 40]$, o conjunto para a quantidade mínima de amostra para divisão sendo $[4, 8, 16, 32, 64, 128]$ e o conjunto para a quantidade minima de amostra para cada folha como $[2, 4, 8, 16]$. De todas as 120 combinações, foi selecionado aleatoriamente 32 valores do *Grid*, particionando no cross validation os dados em 10 amostras e mantendo a SEED igual a 301 para o *Random State* de forma estática para comparar os modelos. Os melhores resultados foram:

```
0.333 +-(0.099) {'min_samples_split': 32, 'min_samples_leaf': 8,
'max_depth': 30, 'criterion': 'gini'}
0.329 +-(0.108) {'min_samples_split': 64, 'min_samples_leaf': 8,
'max_depth': 10, 'criterion': 'gini'}
0.329 +-(0.068) {'min_samples_split': 32, 'min_samples_leaf': 8,
'max_depth': 30, 'criterion': 'entropy'}
```

Para a validação do modelo, foi rodado o cross validation com Fold igual a 10 com acurácia no intervalo de $[22.82\%, 42.25\%]$ e com média 32.53%, o que mostra um resultado superior ao modelo *dummy* que computou-se anteriormente.

```
accuracy medio 32.53
Intervalo [22.82, 42.25]
DecisionTreeClassifier(max_depth=30, min_samples_leaf=8,
min_samples_split=32)
```

Para o modelo de *Machine Learning Random Forest*, utilizou-se o mesmo *Randomized Cross Validation Grid* com um conjunto de valores para a floresta de $[10, 50, 100]$, o conjunto de profundidade máxima como $[3, 5, 10, 15, 20, 30]$, um conjunto para a quantidade mínima de amostra para cada nó como $[4, 8, 16, 32, 64, 128]$ e o conjunto para a quantidade minima de amostra para cada folha como $[4, 8, 16, 32, 64, 128]$. De todas as 540 combinações, selecionou-se aleatoriamente 32 valores do *Grid*, particionando no cross validation os dados em 5 amostras e mantendo a SEED igual a 301 para o *Random State* de forma estática para comparar os modelos. Os melhores resultados foram:

```
0.345 +-(0.038) {'n_estimators': 100, 'min_samples_split': 32,
```

```
'min_samples_leaf': 4, 'max_depth': 20, 'criterion': 'gini',
'bootstrap': True}
0.333 +-(0.051) {'n_estimators': 50, 'min_samples_split': 64,
'min_samples_leaf': 8, 'max_depth': 30, 'criterion': 'entropy',
'bootstrap': False}
0.308 +-(0.060) {'n_estimators': 100, 'min_samples_split': 128,
'min_samples_leaf': 8, 'max_depth': 30, 'criterion': 'gini',
'bootstrap': True}
0.307 +-(0.052) {'n_estimators': 100, 'min_samples_split': 16,
'min_samples_leaf': 8, 'max_depth': 20, 'criterion': 'gini',
'bootstrap': True}
```

O cross validation com Fold igual a 10 ficou com acurácia no intervalo de [26.31%, 43.24%] e com média 34.77%, o que mostra um resultado superior aos modelos *dummy* e *Decision Tree* anteriores.

```
accuracy medio 34.77
Intervalo [26.31, 43.24]
RandomForestClassifier(max_depth=20, min_samples_leaf=4,
min_samples_split=32)
```

Conclusão

Pode-se realizar a seguinte ilação sobre os modelos de machine learning para prever quais as revistas determinado artigo que utiliza o modelo DSGE em macroeconomia deve ser submetido, não somente para ter uma maior margem de aceite, mas também para minimizar a rejeição do artigo que demanda tempo para avaliação.

Dessa maneira, três modelos de *Machine Learning* foram utilizados, um Dummy que serve de *baseline* e outros dois que melhoraram bastante a porcentagem da acurácia. A acurácia com o modelo dummy ficou no intervalo de $16,21\% e 18,43\%$, enquanto que a acurácia do *Decision Tree* e da *Random Forest* ficaram entre $[22.82, 42.25]$ e $[26.31, 43.24]$, respectivamente.

Os DSGEs foram utilizados em larga escala para analisar os impactos da crise do corona virus sobre a economia. Esse trabalho ainda pode ser aperfeiçoado com outros modelos como redes neurais. Isso fica para próximos estudos.

Referências Bibliográficas:

Ambrocio, G., Juselius, M. 2020. Dealing with the costs of the COVID-19 pandemic–what are the fiscal options? BoF Economics Review.

Bar-Ilan, J. 2008. Informetrics at the Beginning of the 21st century-A Review. J. Informetr 2: 1–52. doi:10.1016/j.joi.2007.11.001

Bird, Steven, Loper, E. Klein, E.. 2009. Natural Language Processing with Python. OReilly Media Inc.

Breiman, L. 2001. Random forests. Machine Learning, 45, 5-32.

https://doi.org/10.1023/A:1010933404324.

Campbell, F. 1896. The Theory of National and International Bibliography. With Special Reference to the Introduction of System in the Record of Modern Literature. London: Library Bureau.

Chellappandi, P., Vijayakumar, C. S. 2018. Bibliometrics, Scientometrics, Webometrics/Cybermetrics, Informetrics and Altmetrics - an Emerging Field in Library and Information Science Research. Int. J. Educ. 7, 107–115. doi:10.5281/zenodo.2529398

Eichenbaum, M. S., Rebelo, S., Trabandt, M. 2020. The macroeconomics of epidemics. National Bureau of Economic Research.

Ellegaard, O. 2018. The Application of Bibliometric Analysis: Disciplinary and User Aspects. Sci 1161, 181–202. doi:10.1007/S11192-018-2765-Z

Ellegaard, O., Wallin, J. A. 2015. The Bibliometric Analysis of Scholarly Production: How Great Is the Impact? Scientometrics 105: 1809–1831. doi:10.1007/s11192-015-1645-z

Galí, J., López-Salido, J. D., Vallés, J. 2007. Understanding the effects of government spending on consumption. Journal of the european economic association, 5(1), 227-270 Glänzel, W. 2003. Bibliometrics as a Research Field: A Course on Theory and Application of Bibliometric Indicators. Louven, Belgium: Researchgate.

Godin, B. 2006. On the Origins of Bibliometrics. Scientometrics 68. doi:10.1007/s11192-006-0086-0

González-Alcaide, G. 2021. Bibliometric Studies outside the Information Science and Library Science Field: Uncontainable or Uncontrollable?. Sci 2021, 1–34. doi:10.1007/S11192-021-04061-3

Harzing, A, 2016. Google Scholar, Scopus and the Web of Science: a Longitudinal and Cross-Disciplinary Comparison. Scientometrics 106, 787–804. doi:10.1007/s11192-015-1798-9

Hlavcheva, Y. M., Kanishcheva, O. V., Borysova, N. V. 2019. A Survey of Informetric Methods and Technologies. Cybern Syst. Anal. 55: 503–513doi:10.1007/s10559-019-00158-z

James, G., Hastie T., Tibshirani R. 2013. An Introduction to Statistical Learning. Springer, STS Vol. 103.

Johnson, I. M. 2011. Bibliometrics the Brain Dead. Inf. Dev. 27, 92–93. doi:10.1177/0266666911404012

Jonkers, K., Derrick, G. E. 2012. The Bibliometric Bandwagon: Characteristics of Bibliometric Articles outside the Field Literature. J. Am. Soc. Inf. Sci. Technol. 63, 829–836. doi:10.1002/ASI.22620

Lariviere, V., Sugimoto, C., and Cronin, b. 2012. A Bibliometric Chronicling of Library and Information Science's First Hundred Years. J. Am. Soc. Inf. Sci. Technol. 63, 997–1016. doi:10.1002/asi.22645

Larivière, V. 2012. The Decade of Metrics? Examining the Evolution of Metrics within and outside LIS. Bull. Am. Soc. Inf. Sci. Technol. 38, 12–17. doi:10.1002/BULT.2012.1720380605

McKibbin, W. Fernando, R. 2021. Macroeconomic Policy Adjustments due to COVID-19: Scenarios to 2025 with a focus on Asia. Australian National University;

McKibbin, W. Fernando, R. 2020a. The economic impact of COVID-19. In: BALDWIN, R. DI MAURO, B. (eds.) Economics in the time of COVID-19. London: Centre for Economic Policy Research.

McKibbin, W. Fernando, R. 2020b. The global macroeconomic impacts of COVID-19: seven scenarios. Asian Economic Papers, 19, 1-55.

McKibbin, W. Fernando, R. 2020c. Global macroeconomic scenarios of the COVID-19 pandemic. COVID Economics: Vetted and Real Time papers, 1-58.

Maliszewska, M., Mattoo, A. van der Mensbrugghe, A. 2020. The Potential Impact of COVID-19 on GDP and Trade: A Preliminary Assessment. Policy Research Working Paper. Washington DC: World Bank.

Mooghali, A., Alijani, R., Karami, N., Khasseh, A. 2011. Scientometric Analysis of the Scientometric Literature. Int. J. Inf. Sci. Manag. 9, 19–31.

Pritchard, A. 1969. Statistical Bibliography or Bibliometrics?. J. Doc 25.

Qaiser, S. Ali, R. 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications.

Rose, Michael E. John R. Kitchin. 2010 "pybliometrics: Scriptable bibliometrics using a Python interface to Scopus", SoftwareX 10 100263.

Rousseau, R., Egghe, L., and Guns, R. 2018. Becoming Metric-Wise:A Bibliometric Guide for Researchers - Chapter 1

Rousseau, S., Rousseau, R. 2017. Being metric-wise: Heterogeneity in Bibliometric Knowledge. Prof. La Inf. 26: 480. doi:10.3145/epi.2017.may.14

Schoepflin, U., Glänzel, W. 2001. Two Decades of "Scientometrics" an Interdisciplinary Field Represented by its Leading Journal. Scientometrics 50. doi:10.1023/A:1010577824449

Schubert, A., Glanzel, W., Braun, t. 1989. Scientometric Datafiles - a Comprehensive Set of Indicators on 2649 Journals and 96 Countries in All Major Science fields and Subfields 1981-1985. Scientometrics 16, 3. doi:10.1007/BF02093234

Schubert, A. 2002. The Web of Scientometrics: A Statistical Overview of the First 50 Volumes of the Journal. Scientometrics 53. doi:10.1023/A:1014886202425

World Bank. 2020a. East Asia and Pacific Economic Update: East Asia and Pacific in the time of COVID-19. Washington DC: World Bank.

World Trade Organization. 2020. Methodology for the WTO Trade Forecast as of April 2 [Online]. Geneva: Economic Research Statistics Division, World Trade Organization. Available: https://www.wto.org/english/news_e/pres20_e/methodpr855_e.pdf

Zhang, Y., Zhang, G., Zhu, D., Lu, J. 2017. Scientific Evolutionary Pathways: Identifying and Visualizing Relationships for Scientific Topics. J. Assoc. Inf. Sci. Technol. 68, 1925–1939. doi:10.1002/asi.23814

Zuccala, A. 2016. Inciting the Metric Oriented Humanist: Teaching Bibliometrics in a Faculty of Humanities. Educ. Inf. 32: 149–164. doi:10.3233/EFI-150969

Zhang, X., Zhang, Y., Zhu, Y. 2021. COVID-19 Pandemic, Sustainability of Macroeconomy, and Choice of Monetary Policy Targets: A NK-DSGE Analysis Based on China. Sustainability 2021, 13, 3362. https://doi.org/10.3390/su13063362

Apêndice A - Montando a Base de Dados e realizando o préprocessamento

Pacotes Pyhthon

```
import numpy as np
import pandas as pd
import spacy

import seaborn as sns
import matplotlib.pyplot as plt

from pybliometrics.scopus import ScopusSearch , CitationOverview ,
AbstractRetrieval

from spacy.lang.en.stop_words import STOP_WORDS as stopwords

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.dummy import DummyClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import GroupKFold , RandomizedSearchCV ,
KFold , cross_validate , cross_val_score
```

Pré-processamento da base de dados

```
df['descricao_proc'] = df['descricao'].apply(lambda x: cont_to_exp(x))

#removendo chracter special
df['descricao_proc'] = df['descricao_proc'].apply(lambda x: re.sub(r'
[^\w ]+', "", x))
```

Selecionando as revistas como variáveis explicativas categóricas

```
dsge_filter = df.loc[df['descricao_proc'].str.contains('monetary') &
df['ano'].str.startswith('20')]
.;# Filtrando por 2 parametros
len (dsge_filter)

veiculos=['economic modelling', 'journal economic dynamic
control','journal money , credit banking','journal macroeconomic',
'journal monetary economic','macroeconomic dynamic','international
journal central banking' ,'journal international money finance']
```

Apêndice B - Código utilizado para os modelos de Machine Learning

Aplicando o TFIDF

```
tfidf = TfidfVectorizer()
x = tfidf.fit_transform(dsge_filter['descricao_proc'])

y = dsge_filter['veiculo']
```

Aplicando o modelo Dummy

```
SEED = 301
np.random.seed(SEED)

modelo = DummyClassifier()
results = cross_validate(modelo, x, y, cv = 10, return_train_score
=False)
media = results['test_score'].mean()
desvio_padrao = results['test_score'].std()
print("Accuracy com dummy stratified, 10 = [%.2f, %.2f]" %
((media - 2 * desvio_padrao)*100, (media + 2 * desvio_padrao) * 100))
```

O modelo de de Floresta Aleatória

```
SEED=301
np.random.seed(SEED)

espaco_de_parametros = {
    "n_estimators" : [10, 50, 100],
    "max_depth" : [3, 5, 10, 15, 20, 30],
    "min_samples_split" : [4, 8, 16, 32, 64, 128],
    "min_samples_leaf" : [4, 8, 16, 32, 64, 128],
    "bootstrap" : [True, False],
    "criterion" : ["gini", "entropy"]
}

busca = RandomizedSearchCV(RandomForestClassifier(),
                    espaco_de_parametros,
                    n_iter = 32,
                    cv = KFold(n_splits = 5, shuffle=True))
busca.fit(x, y)

resultados = pd.DataFrame(busca.cv_results_)
resultados.head()

resultados_ordenados_pela_media =
resultados.sort_values("mean_test_score", ascending=False)
for indice, linha in resultados_ordenados_pela_media[:5].iterrows():
```

```
  print("%.3f +-(%.3f) %s" % (linha.mean_test_score,
  linha.std_test_score*2, linha.params))

scores = cross_val_score(busca, x, y, cv = KFold(n_splits=5,
shuffle=True))
imprime_score(scores)

melhor = busca.best_estimator_
print(melhor)
```

O modelo de de Arvore de Decisao

:

```
SEED=301
np.random.seed(SEED)
espaco_de_parametros = {
    "max_depth" : [10, 15, 20, 30, 40 ],
    "min_samples_split" : [4, 8, 16, 32, 64, 128],
    "min_samples_leaf" : [2, 4, 8, 16],
    "criterion" : ["gini", "entropy"]
}

busca = RandomizedSearchCV(DecisionTreeClassifier(),
                       espaco_de_parametros,
                       n_iter = 32,
                       cv = KFold(n_splits = 10, shuffle=True),
                       random_state = SEED)
busca.fit(x, y)
resultados = pd.DataFrame(busca.cv_results_)
resultados.head()

resultados_ordenados_pela_media =
resultados.sort_values("mean_test_score", ascending=False)
for indice, linha in resultados_ordenados_pela_media.iterrows():
  print("%.3f +-(%.3f) %s" % (linha.mean_test_score,
  linha.std_test_score*2, linha.params))


scores = cross_val_score(busca, x, y, cv = KFold(n_splits=5,
shuffle=True))
imprime_score(scores)
melhor = busca.best_estimator_
print(melhor)
```