**RECONHECIMENTO DE PADRÕES DE INTERESSE EM DADOS FISCAIS UTILIZANDO APRENDIZAGEM DE MÁQUINA NÃO SUPERVISIONADA**

**Gabriel Spinola Magnavita1**; Ilan Sousa Figueiredo2, Wenisten Jose Dantas da Silva2, Lílian Lefol Nani Guarieiro2, Erick Giovani Sperandio Nascimento2

1Bolsista IC voluntário; Tipo de projeto; gabriel.magnavita@fbter.org.br

2 Centro Universitário SENAI CIMATEC; Salvador-BA; erick.sperandio@fieb.org.br

**RESUMO**

O presente trabalho utilizou um algoritmo baseado em aprendizado de máquina não supervisionada para reconhecer padrões de interesse na economia brasileira. O banco de dados aplicado consistiu na quantidade de notas fiscais (NF) emitidas por 12 estados brasileiros desde 01 de janeiro de 2019 a 08 de fevereiro de 2021. O algoritmo foi capaz de reconhecer três padrões de interesse: (i) normal, (ii) *lockdown* ocasionado pela COVID-19 e (iii) período de festas e celebrações de fim de ano. Os resultados demonstraram que o algoritmo C-AMDATS foi capaz de reconhecer esses padrões com alta performance. Os resultados revelaram o seguinte desempenho: Acurácia de 93%, Precisão de 79%, Sensibilidade de 85%, Pontuação F1 de 82%, Especificidade de 85%, Curva AUC-ROC de 90% e Curva AUC-PRC de 76%. Portanto, concluímos que a aprendizagem de máquina não supervisionada é capaz de automaticamente reconhecer padrões de interesse na economia brasileira, revelando grande potencial em revelar informações ocultas em grandes volumes de dados sem a necessidade de uma inspeção visual por especialistas.

**PALAVRAS-CHAVE:** reconhecimento de padrões, aprendizado de máquina não supervisionada, séries temporais multivariadas, dados fiscais.

**1. INTRODUÇÃO**

Atualmente, com a crescente proliferação do SARS-CoV-2 (COVID-19) em escala global e regional, pesquisadores e agentes públicos trabalham conjuntamente para estabelecer uma série de medidas de proteção pública que visa mitigar a proliferação do SARS-CoV-2. Essas medidas de mitigação consistem principalmente em reduzir o número de pessoas infectadas e o número de óbitos decorrentes do vírus.

A título de exemplo, Chaudhary e Singh (2020) analisaram a tendência de 187 países afetados regionalmente e a variação dos casos no nível de cada país.1 O experimento usou a análise de componentes principais (PCA) das variáveis do conjunto de dados COVID-19 para reduzir a dimensionalidade e encontrar as variáveis mais significativas. Em seguida, usaram a aprendizagem de máquina não supervisionada k-means com o método Elbow (identifica o número ideal de agrupamentos) para revelar os padrões ocultos dos países afetados pelo SARS-CoV-2. Os resultados revelaram 6 e 7 agrupamentos entre os países. A pesquisa sugere que o experimento poderia contribuir à formulação de várias políticas de segurança no setor da saúde e ajudar médicos e especialistas em economia. Além disso, o estudo poderia ser útil para revelar os países mais impactados pelo vírus, de modo, a ajudá-los a receber assistência dos menos impactados, e auxiliar governantes na tomada de ações preventivas.1

Essas políticas quando são adotadas geram impactos tanto na área de saúde, evitando superlotação dos leitos de hospitais, como na economia, menos circulação de capital. Ao adotar medidas de restrição, como o fechamento do comércio por exemplo, é esperado que ocorra alterações no número de emissões de NF devido a suspensão das atividades não essenciais, assim como espera-se alterações nesses dados no período de fim de ano devido ao décimo terceiro salário e celebrações (*ex.*: natal e réveillon).

Com base nessa informação, este trabalho avalia o desempenho do algoritmo de aprendizagem de máquina não supervisionada C-AMDATS a reconhecer padrões de interesse em dados de série temporal da quantidade de NF emitidas por 12 estados brasileiros. Essa abordagem permite descobrir informações ocultas em grandes volumes de dados sem a necessidade de uma inspeção visual por especialistas.2

**2. METODOLOGIA**

Foi utilizado dados disponibilizados pela SEFAZ do Rio Grande do Sul a quantidade das NF emitidas no Brasil no período de 01 de janeiro de 2019 a 08 de fevereiro de 2021.3 Esses dados foram processados (dados brutos) e analisados diretamente no algoritmo C-AMDATS para reconhecimento de padrões de interesse dos períodos de *lockdown* ocasionado pela COVID-19 e das celebrações de fim de ano.

Para definir os valores desses períodos de interesse, fez-se uma análise analítica das medidas de mitigação adotadas por 12 estados brasileiros e, para o período fim de ano, foi considerado uma janela de tempo de 30 dias antes do Natal. A Tabela 1 exibe as medidas de mitigação que foram consideradas nessa análise.

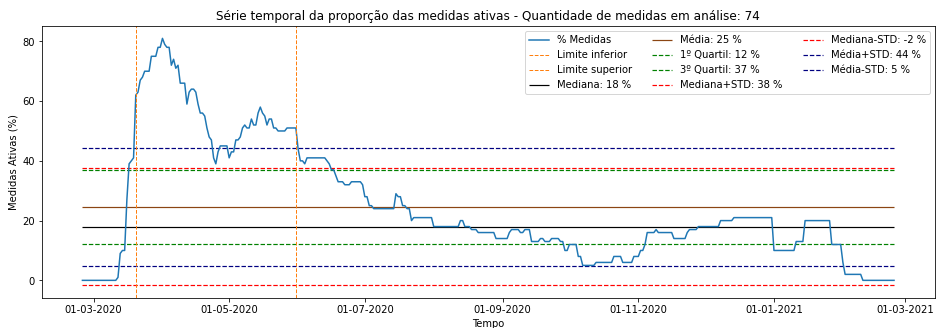
Tabela 1. Medidas de mitigação consideradas na análise.

|  |
| --- |
| Medida Descrição |
| C1 Fechamento da escola |
| C2 Fechamento do local de trabalho |
| C3 Cancelar eventos públicos |
| C4 Restrições ao tamanho da coleta |
| C5 Transporte público próximo |
| C6 Ordens de “abrigo no local” e confinamento domiciliar |
| C7 Restrições ao movimento interno |
| C8 Restrições a viagens internacionais |

O levantamento das medidas de mitigação foi realizado de forma binária, isto é, ativa (1) ou inativa (0).

A Figura 1 ilustra a série temporal da porcentagem das medidas ativas de forma simultânea. Portanto, através dessa informação foi possível revelar a janela de tempo que apresentou maior concentração de medidas ativas da COVID-19 no Brasil.

Figura 1. Visualização da série temporal das medidas adotadas pelos estados brasileiros de forma síncrona durante o período de 01/03/2020 até 01/03/2021.



A partir da curva das medidas ativas (Figura 1), alguns indicadores estatísticos como média, mediana, 1° quartil, 3° quartil e desvio padrão foram calculados para definir o período de interesse, ou seja, período de restrição máxima adotado pelos governantes durante a pandemia. A variação da média com o desvio padrão foi adotada como limiar para revelar o período de *lockdown*. Portanto, todos os valores do eixo y acima da média mais o desvio padrão () foram considerados como período de *lockdown* máximo. A Tabela 2 exibe os limiares inferiores e superiores do período de *lockdown* identificado na Figura 1.

Tabela 2. Limite inferior e superior do período de *lockdown* ocasionado pela COVID-19

|  |
| --- |
| Indicador Data Medida (%) |
| Limite Inferior 20-02-2020 62 |
| Limite Superior 31-05-2020 51 |
| Valor Máximo 01-04-2020 81 |

Portanto, com base nas análises realizadas nos dados das medidas de mitigação, foi possível definir o período de interesse em que o algoritmo de aprendizagem de máquina não supervisionada deverá reconhecer de forma automática, além também do período de festas e comemorações do fim de ano (25 de novembro a 25 de dezembro).

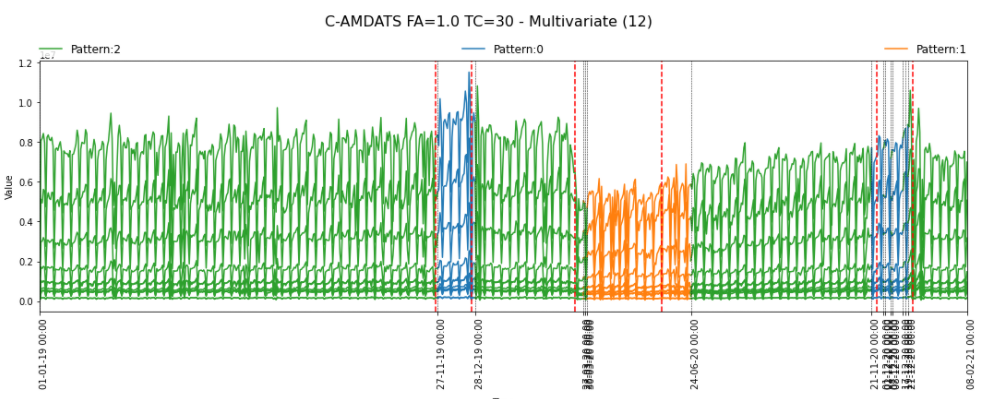
O algoritmo selecionado para fazer as análises foi o C-AMDATS, pois esse algoritmo de aprendizagem de máquina não supervisionada tem revelado alto desempenho a várias outras técnicas, principalmente devido a sua abordagem multivariada e agrupamento através da distância de Mahalanobis.4 Além disso, o algoritmo é simples de usar, possuindo apenas dois hiperparâmetros (Tamanho Inicial de Agrupamento e Fator de Agrupamento) a serem definidos pelo usuário. Maiores informações sobre o C-AMDATS podem ser encontradas em Nascimento et al., (2015).2

**3. RESULTADOS E DISCUSSÃO**

O algoritmo C-AMDATS conseguiu de forma inteligente e não supervisionada reconhecer os padrões de interesse propostos neste estudo. Foram utilizados os valores de 30 dias e 1.0 nos parâmetros Tamanho Inicial de Agrupamento e Fator de Agrupamento, respectivamente.

A Figura 2 ilustra os padrões reconhecidos pelo C-AMDATS na série temporal da quantidade de NF emitidas por 12 estados brasileiros.

Figura 2. Visualização do reconhecimento de padrões do C-AMDATS na série temporal multivariada durante o período de 01/03/2020 até 01/03/2021.



Os padrões 0 e 1 foram considerados como anômalos e 2 como normal. O padrão 0 consistiu no padrão de fim de ano quando ocorre um aumento nas compras em decorrência dos assalariados receberem o décimo terceiro salário, ocorrendo um aumento significativo na quantidade de emissão de NF na economia. O padrão 1 foi o período da primeira onda da COVID-19. As duas linhas verticais tracejadas em vermelho são as datas escolhidas como início e fim dos valoresdos padrões de interesse.

O algoritmo C-AMDATS reconheceu os padrões de interesse com alta performance. Portanto, concluímos que a nossa análise não supervisionada é capaz de automaticamente reconhecer eficientemente distintos padrões na economia brasileira.

**4. CONSIDERAÇÕES FINAIS**

Neste trabalho foi utilizado o algoritmo C-AMDATS, de aprendizagem de máquina não supervisionada, para detectar padrões de interesse em dados de séries temporais da quantidade de NF emitidas em 12 estados brasileiros. Os padrões de interesse concerniram aos períodos de *lockdown* ocasionado pela COVID-19 e festas comemorativas de fim de ano. O algoritmo foi executado em dados reais e obteve uma Acurácia de 93%, Precisão de 79%, Sensibilidade 85%, Pontuação F1 82%, Especificidade 85%, Curva AUC ROC 90%, Curva AUC PRC 76%. O método permite detectar padrões de anomalias de forma eficiente em séries temporais multivariadas e é útil para extrair informações ocultas em grande volume de dados.

**Agradecimentos**

Agradecemos ao Centro de Referência em Inteligência Artificial, além do Centro de supercomputação para Inovação Industrial (CS2i), ambos do SENAI CIMATEC. Agradecemos também ao SENAI CIMATEC pelo financiamento desta pesquisa.

**5. REFERÊNCIAS**

1 CHAUDHARY, Laxmi; SINGH, Buddha. Community Detection using Unsupervised machine learning technique on COVID-19 dataset. 2020.

2 Nascimento, E. G. S., De Lira Tavares, O. & De Souza, A. F. A cluster-based algorithm for anomaly detection in time series using mahalanobis distance. Proc. 2015 Int. Conf. Artif. Intell. ICAI 2015 - WORLDCOMP 2015 622–628 (2015).

3 Banco de dados da Receita Estadual. Receita Dados, 2021. Dados Abertos. Disponível em: <http://receitadados.fazenda.rs.gov.br/Arquivos/Valor%20e%20quantidade%20de%20DFe%20por%20dia.csv>. Acesso em: 06 de abr. de 2021.

4 FIGUEIRÊDO, Ilan; GUARIEIRO, Lílian Lefol Nani; NASCIMENTO, Erick Giovani Sperandio. Multivariate Real Time Series Data Using Six Unsupervised Machine Learning Algorithms. In: **Anomaly Detection-Recent Advances, Issues and Challenges**. IntechOpen, 2020.