Evento: XXVII Jornada de Pesquisa

PROPOSTA DE DESENVOLVIMENTO DE UM MODELO COMPUTACIONAL BASEADO EM MACHINE LEARNING PARA ENCONTRAR CORRELAÇÕES ENTRE O PERFIL ECONÔMICO E O NÍVEL DE ISOLAMENTO SOCIAL NO CONTEXTO DA PANDEMIA DE COVID-19 NO ESTADO DO RIO GRANDE DO SUL[1]

PROPOSAL FOR THE DEVELOPMENT OF A COMPUTATIONAL MODEL BASED ON MACHINE LEARNING TO FIND CORRELATIONS BETWEEN THE ECONOMIC PROFILE AND THE LEVEL OF SOCIAL ISOLATION IN THE CONTEXT OF THE COVID-19 PANDEMIC IN THE STATE OF RIO GRANDE DO SUL

Félix Hoffmann Sebastiany[2], Matheus Henrique Rehbein[3], Rafael Z. Frantz[4], Sandro Sawicki[4]

[1] Projeto de pesquisa desenvolvido no PPGMMC/Unijuí no âmbito do Grupo de Pesquisa em Computação Aplicada (GCA)

[2] Mestrando em Modelagem Matemática e Computacional na Unijuí, bolsista CAPES
[3] Doutorando em Modelagem Matemática e Computacional na Unijuí. bolsista CAPES
[4] Professor Doutor do Programa de Pós-Graduação em Modelagem Matemática e Computacional da Unijuí

RESUMO

Em 2020, a pandemia de Covid-19 afetou diferentes pessoas de várias maneiras. A maioria das pessoas infectadas apresentaram sintomas leves a moderados da doença, entretanto caso não haja medidas protetivas, a contaminação desenfreada pode gerar sobrecarga nos sistemas hospitalares. No Brasil, assim como em grande parte de outros países, adotou-se o isolamento social como forma de controlar a propagação do vírus. Os índices de isolamento diário foram fornecidos aos gestores públicos para auxiliar no controle da pandemia. No entanto, muitas dúvidas ainda estão presentes quanto à eficácia desse método. Nesse sentido, mais análises são necessárias para compreender e interpretar de forma mais efetiva a relação entre isolamento social e os novos casos de contaminação. Neste contexto, este trabalho busca utilizar a base de dados CNAE (Classificação Nacional de Atividades Econômicas), juntamente com as bases de dados de Isolamento Social (InLoco) e número de casos diários de Covid-19 (Secretaria de Saúde do Estado do Rio Grande do Sul) com o objetivo de desenvolver um modelo computacional para descoberta de conhecimento a partir da relação entre essas bases de dados no contexto da pandemia de Covid-19 no Estado do Rio Grande do Sul usando *Machine Learning*.

Palavras-chave: Aprendizado de Máquina. Inteligência Artificial. Covid-19. Algoritmos de Clusterização

ABSTRACT
In 2020, the Covid-19 pandemic affected different people in different ways. Most infected people with symptoms may have disease moderators, take protective measures, cases of unrestrained hospital illness generate in the systems. In Brazil, as in most other countries, as a form of control, it is controlled by the social virus. Daily isolation rates were provided to public managers to help control the pandemic. However, many doubts are still present regarding the technique of this method. In this sense, more considered are considered to interpret and more the between social isolation and the new cases of consideration. In this case, using a CNAE database (InLoco) and Covid-19 data number from the Rio State Health Department (Daily Work (CNAE Class) do Sul) with the aim of developing a computational model for discovery of knowledge from the relationship between these databases in the context of the Covid-19 pandemic in the State of Rio Grande do Sul using Machine Learning.

Keywords: Machine Learning. Artificial Intelligence. Covid-19, Clustering Algorithms.

INTRODUÇÃO

Machine Learning, ou, no português, Aprendizado de Máquina, é um método de análise de dados que proporciona a automatização da construção de modelos analíticos, o qual inclui técnicas estatísticas para aprender por meio da experiência e do treinamento. É uma sub-área da Inteligência Artificial com base na premissa de que sistemas podem aprender com dados por meio da identificação de padrões e tomar decisões assertivas com o mínimo de intervenção humana.

No final do ano de 2019 ocorreu na cidade de Wuhan / China, um surto de uma nova doença respiratória provocada pelo novo coronavírus, intitulada Covid-19. Em março de 2020, a Organização Mundial da Saúde (OMS) declarou situação de pandemia mundial em decorrência deste fato.

Uma das estratégias para o enfrentamento desta pandemia é o isolamento das pessoas contaminadas, pois representam risco de contágio para outras pessoas. Esta prática envolve, também, medidas de distanciamento social entre as pessoas aparentemente sadias com vistas a impedir o colapso do sistema de saúde. O isolamento social tem sido adotado pelo Brasil e por vários países durante a pandemia na tentativa de atenuar a curva de contágio da doença. O debate sobre a importância do isolamento social e de outras medidas voltadas à restrição da circulação de pessoas nas cidades vem sendo motivado por vários segmentos da sociedade.

No Brasil, entretanto, isso vem gerando confusão e dúvidas sobre a eficácia desse método e quais ações devem ser adotadas para segurar o contágio pelo Covid-19.

Neste contexto, este trabalho busca utilizar a base de dados CNAE (Classificação Nacional de Atividades Econômicas) juntamente com as bases de dados de Isolamento Social (InLoco) e número de casos diários de Covid-19 (Secretaria de Saúde do Estado do Rio Grande do Sul) com o objetivo de desenvolver um modelo computacional para descoberta de conhecimento a partir da relação entre essas bases de dados no contexto da pandemia de Covid-19 no Estado do Rio Grande do Sul usando *Machine Learning*.

METODOLOGIA

Inicialmente, foi realizado um estudo do estado-da-arte sobre Inteligência Artificial visando compreender suas características, suas classificações, bem como seus modos de aplicações através de livros e artigos científicos.

Por meio da análise da base de dados CNAE (Classificação Nacional de Atividades Econômicas) busca-se definir o critério de agrupamento com vistas a correlacionar grupos com similaridade econômica, além de realizar sua caracterização e detalhamento. Após, tabular os grupos e cruzar com as médias proporcionais de isolamento social em um determinado intervalo de tempo.

Com base no vínculo empregatício de cada município, o modelo de clusterização irá atribuí-los para cada grupo criado na análise dos dados do CNAE. Concomitantemente, será implementado algoritmos para interpretar os dados de isolamento social por municípios disponibilizados pela empresa InLoco e dos dados de casos diários de Covid-19 por municípios do RS disponibilizados pela Secretaria da Saúde do Estado do Rio Grande do Sul.

Com a estrutura de dados criada, será desenvolvido um modelo computacional com o objetivo de cruzar as informações dos perfis socioeconômicos, dos dados de isolamento social e dos casos diários de Covid-19 dos municípios do Estado do Rio Grande do Sul com o objetivo de encontrar padrões e relacionamentos. Para a implementação desse modelo computacional será utilizada a linguagem de programação *Python*.

Esta pesquisa classifica-se em exploratória explicativa. Exploratória de modo a investigar o conhecimento acerca dos dados brasileiros de Covid-19, bem como compreender os grupos socioeconômicos propostos pelo CNAE e o isolamento social. E explicativa através da utilização de dados reais como subsídio para a obtenção de um modelo de Inteligência Artificial representativo da relação entre casos de Covid-19 com grupos socioeconômicos.

Quanto à fonte de pesquisa, considera-se como secundária, devido a utilização de dados experimentais coletados através de bases consolidadas. Com isto, objetiva-se obter resultados quantitativos provenientes das análises feitas através do modelo de *Machine Learning*.

PROPOSTA

Para reduzir as altas taxas de transmissão do vírus, a OMS (Organização Mundial da Saúde) propôs o isolamento social como prática no cotidiano da sociedade. Essa medida pode ou não ser seguida e depende de fatores como consentimento social e político. No entanto, ainda surgem dúvidas sobre a eficácia do isolamento social no combate à disseminação do vírus.

Este projeto busca correlacionar as seguintes bases de dados: CNAE (Classificação Nacional de Atividades Econômicas), Isolamento Social (InLoco) e da Secretaria Estadual de Saúde do Estado do Rio Grande do Sul, a qual disponibiliza dados diários de contaminação e óbitos. A partir dessa correlação, por meio de um algoritmo de *Machine Learning*, temos como proposta do trabalho, responder algumas hipóteses que foram levantadas. As hipóteses são:

- Hipótese 1: Cidades com perfil econômico no setor secundário (indústria) tiveram um número maior de contágio por Covid-19;
- Hipótese 2: Cidades em que a atividade econômica é predominantemente comercial (setor terciário - bens e prestação de serviços) tiveram maior número de contágio por Covid-19;
- Hipótese 3: Cidades com perfil na agricultura (setor primário) tiveram menos casos de contaminação;

EXPERIMENTAÇÃO

1. ELABORAÇÃO DE *DATASET*

Para a realização desta pesquisa são utilizadas no experimento três principais bases de dados. A primeira refere-se aos dados socioeconômicos de cada município do estado do Rio Grande do Sul. A segunda fornece dados de casos resultantes do Covid-19. A terceira fornece índices de isolamento social diário dos municípios do estado do Rio Grande do Sul. Através dessas bases de dados foi criado um *dataset* final, detalhado na Seção 1.4.

Todos os dados foram tratados utilizando a linguagem de programação *Python*, que é uma linguagem de código aberto amplamente utilizada em *data science*.

1.1. CNAE - Classificação Nacional de Atividades Econômicas

É obrigatório que todas as pessoas jurídicas, inclusive autônomos e organizações sem fins lucrativos, a CNAE é essencial para obtenção do CNPJ (Cadastro Nacional da Pessoa Jurídica). Além de contribuir para melhorar a gestão tributária do país, também é possível fazer um levantamento de qual a atividade econômica mais exercida em um determinado município, assim como também é possível saber os vínculos empregatícios existentes.

O CNAE divide-se em 21 grupos de atividades econômicas. Rehbein (2020), propôs um algoritmo de clusterização chamado *K-Means*, que divide os grupos em 7, resultando em um maior agrupamento de dados. Considerando esse algoritmo, o *dataset* do CNAE, possui uma lista de todos os municípios do RS, com sua principal atividade econômica, entre as 7:

- Classe 0: Administração Pública e Comércio e Atividades Essenciais;
- Classe 1: Transformação / Industrial;
- Classe 2: Agropecuário e Administração Pública;
- Classe 3: Transformação / Industrial e Administração Pública;
- Classe 4: Transformação / Industrial e Comércio e Atividades Essenciais;
- Classe 5: Administração Pública;

- Classe 6: Comércio e Atividades Essenciais e Administração Pública.

1.2. Secretaria de Saúde do Estado do Rio Grande do Sul

Para a pesquisa, foram coletados dados do número de casos confirmados de Covid-19 diariamente em todos os municípios do estado. A Secretaria de Saúde do Estado do Rio Grande do Sul, disponibiliza esses dados através do "Painel Coronavírus RS", criado para apresentar os principais dados epidemiológicos da Covid-19 no RS.

1.3. InLoco

A empresa InLoco, que realiza a apresentação dos Índices de Isolamento Social, fornece dados referentes à aceitação e cumprimento das restrições voltadas ao isolamento durante a pandemia, servindo como fonte de referências para pesquisas e até mesmo para os órgão públicos, segundo Moura e Ferraz (2020). Toda a coleta de dados foi realizada com autorização dos participantes, autorizando a coleta de dados, através da instalação de aplicativos parceiros da empresa InLoco.

Moura e Ferraz (2020), também explicam que diante do uso de tecnologias de geolocalização, realizava-se o rastreamento das pessoas na busca pela propagação do vírus, alimentando um mapa que permite a verificação do índice de isolamento social. Até as compras realizadas em lojas físicas onde os cartões de crédito eram a forma de pagamento, servia de dados para alimentação do mapa ao cruzar as informações com a localização da residência do indivíduo, permitindo traçar uma rota percorrida. Assim, diversos países passaram a adotar medidas de controle de geolocalização no combate ao Covid-19. A proposta da empresa InLoco foi bem aceita e dentro das questões de privacidade, apresentava uma solução ao governo, prefeituras e pesquisadores, diante da integração de seus *softwares* aos aplicativos do poder público, com métricas de isolamento social por região indicação de áreas de risco, índice de deslocamento de pessoas, visita a estabelecimentos, pontos de aglomeração e através da comunicação com a população.

Porém, em março de 2021, a empresa teve sua prestação de conteúdo descontinuada devido às determinações da Lei Geral de Proteção de Dados, que passaram a influenciar na

coleta de informações. Em sua última coleta de dados, realizada em março de 2021, o estado do Rio Grande do Sul apresentou um índice de isolamento social de 36,2%.

O *dataset* gerado a partir dos dados coletados, apresenta dados diários de isolamento social de cada um dos municípios do estado do Rio Grande do Sul, do dia 17/04/2020 até 31/03/2021.

1.4. *Dataset* Final

O *dataset* final, que é o *dataset* utilizado no algoritmo de clusterização dessa pesquisa, contempla um *merge* entre as três bases de dados citadas. O *dataset* final é composto por uma tabela de 8179 linhas, a qual possui agrupamentos semanais dos dados requeridos: número de casos confirmados de Covid-19, classificação econômica e isolamento social de todos os municípios do estado do Rio Grande do Sul.

A Tabela 1 apresenta alguns dados retirados do *dataset* final.

Tabela 1 - Prévia do *dataset* final, contendo informações semanais de isolamento social e casos de covid de cada município do estado do RS.

Semana	Cidade	Classificação Econômica de predominância	Média de Isolamento Social na semana	Média de Casos por dia Confirmados na semana
2020-06-01/ 2020-06-07	Ijuí	Comércio e Atividades Essenciais e Administração Pública	40%	22,8
2020-12-21/ 2020-12-27	Ijuí	Comércio e Atividades Essenciais e Administração Pública	40%	3184,4
2020-06-01/ 2020-06-07	Santa Rosa	Administração Pública e Comércio e Atividades Essenciais	39%	22,2
2020-12-21/ 2020-12-27	Santa Rosa	Administração Pública e Comércio e Atividades Essenciais	36%	3856,7
2020-06-01/ 2020-06-07	Panambi	Agropecuário e Administração Pública	41%	1,14
…	…	…	…	…

2. ALGORITMO DE CLUSTERIZAÇÃO

Com o objetivo de correlacionar os dados descritos no *dataset* e encontrar padrões que positivem ou não as hipóteses, foi criado um algoritmo de clusterização. Clusterização é uma técnica de *Machine Learning* (Inteligência Artificial), de aprendizado não supervisionado que clusteriza os dados conforme suas próprias características. Existem vários algoritmos de clusterização, entre eles, o *K-Means*, que foi o algoritmo escolhido para essa pesquisa.

A linguagem de programação *Python* foi escolhida para o desenvolvimento do algoritmo, visto que é uma linguagem amplamente utilizada pela comunidade de *data science* e possui várias bibliotecas úteis para o desenvolvimento da pesquisa. Entre as principais bibliotecas utilizadas estão *pandas*, *scikit-learn*, *plotly*. *Pandas* é uma biblioteca criada para manipulação e análise de dados. *Scikit-learn* é uma biblioteca desenvolvida especificamente para aplicação prática de *machine learning*, a qual disponibiliza, por exemplo, o *K-Means*. *Plotly* é uma biblioteca para visualização de dados através de gráficos interativos. Todas as bibliotecas utilizadas são de código aberto.

O algoritmo *K-Means* realiza o agrupamento de dados de acordo com suas características e semelhanças. Em sua implementação deve-se definir o número de *clusters* (agrupamentos) que desejamos gerar. Uma técnica amplamente utilizada para definir o número de *clusters* ideais a serem gerados, é a utilização do Coeficiente de Silhueta.

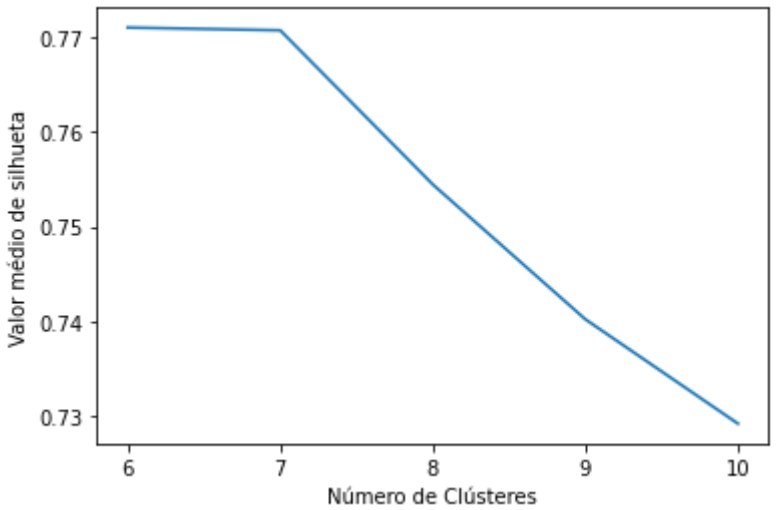
No Coeficiente de Silhueta é analisado um coeficiente resultante de um cálculo de distância entre os centróides levando em consideração o agrupamento dos dados que os cerca. Após o cálculo é gerado um gráfico em barras horizontais que mostra em cada *cluster* o valor do coeficiente dos dados mais próximos aos mais distantes, formando assim uma “silhueta”. O coeficiente é gerado entre o intervalo de -1 a +1 e quando se aproxima ao +1, representa maior distância entre os *clusters* representando um melhor agrupamento. O Código 1 apresenta a implementação do Coeficiente de Silhueta.

Código 1 - Implementação do coeficiente de silhueta para o *dataset* final.

```
### AVALIAÇÃO DE CLUSTERS - coeficiente de silhueta K-MEANS
faixa_n_clusters = [i for i in range(6,13)]
valores_silhueta = []
for k in faixa_n_clusters:
   agrupador = KMeans(n_clusters=k)
   labels = agrupador.fit_predict(final_dataset_semanas)
   media_silhueta = silhouette_score(final_dataset_semanas, labels)
   valores_silhueta.append(media_silhueta)
```

A Figura 1 apresenta um gráfico gerado com os valores médios de silhueta para um *range* de *clusters*. Através da figura, observa-se uma média de silhueta satisfatória quando utilizamos 7 *clusters*. Considerando isso, serão utilizados 7 *clusters* no algoritmo *K-Means*.

Figura 1 - Valores médios de silhueta entre os *clusters* 6 ao 10.



Antes de realizar a clusterização, é importante normalizar os dados do *dataset*. A normalização de dados é uma prática típica em *machine learning*, que consiste em uma técnica de preparação de dados com objetivo de mudar os valores das colunas numéricas no conjunto de dados para usar uma escala comum, sem distorcer suas diferenças. Considerando isso, foi aplicado a função *min-max* da biblioteca *Scikit-learn*, que transforma os valores

numéricos em uma escala entre 0 e 1. O Código 2 apresenta a implementação da normalização de dados, utilizando *Scikit-learn* e *Pandas*.

Código 2 - Implementação da normalização de dados para o *dataset* final.

```
### NORMALIZAÇÃO DOS DADOS
min_max_scaler = preprocessing.MinMaxScaler();
np_df = min_max_scaler.fit_transform(final_dataset_semanas);
df = pd.DataFrame(np_df, columns = final_dataset_semanas.columns)
```

Finalmente, com o número de *clusters* definidos e o *dataset* normalizado, podemos implementar o *K-Means*, para criar os agrupamentos. O código 3 apresenta essa implementação. A primeira linha, invoca a biblioteca *Scikit-learn* (*lib_cluster*), e define que será utilizado o algoritmo *K-Means*, utilizando 7 *clusters* e aplicando o método *fit()* ao *dataset* final (df). O método *fit()* é utilizado para treinar os dados. O parâmetro adicionado ao método *fit* é nosso *dataset* final, já tratado e normalizado, contendo apenas os dados que queremos realizar o agrupamento: índice de isolamento social, média de casos de Covid-19 e classe de atividade econômica. O método *.labels_* implementado na segunda linha do código, retorna a qual *cluster* gerado pertence cada uma das linhas do *dataset* final, e isso é atribuído a uma nova coluna do *dataset* final, que será utilizada na visualização de dados.

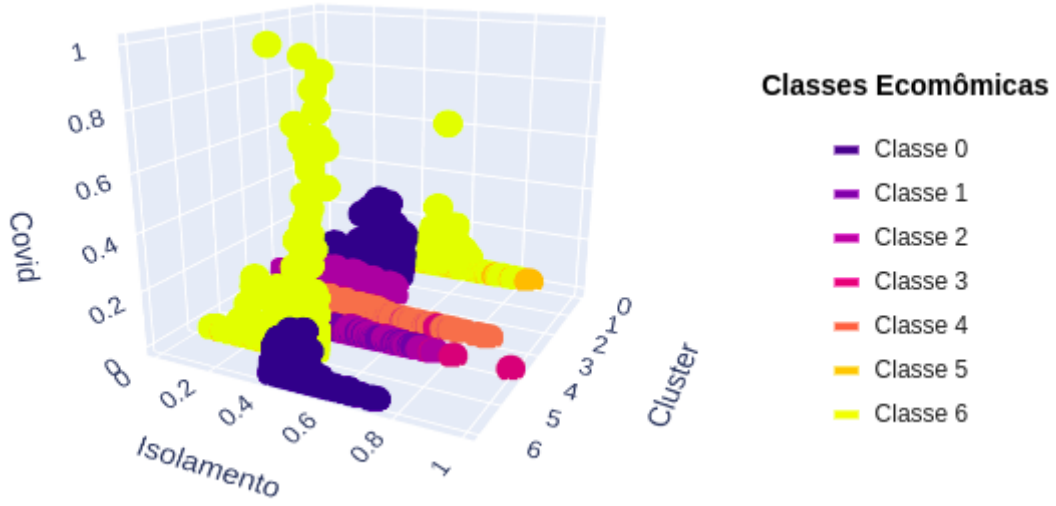
Código 3 - Implementação da clusterização aplicada ao *dataset* final.

```
#### CLUSTERING COM KMEANS
cluster = lib_cluster.KMeans(n_clusters=7).fit(df)
df['cluster'] = cluster.labels_;
```

3. VISUALIZAÇÃO E INTERPRETAÇÃO DE DADOS

Para visualizar e interpretar os dados gerados pelo modelo de clusterização, utilizou-se a biblioteca *Plotly*. Através dela foi gerado um gráfico de 3 dimensões, que é apresentado na Figura 2.

Figura 2 - Agrupamento da correlação entre dados de Isolamento Social, Classe Econômica e Índice de casos de Covid-19 dos municípios do estado do Rio Grande do Sul.



Os dados contidos na Figura 2 mostram que existe uma tendência maior no índice de Covid-19 quando se tem menos isolamento social. Os *clusters* com maior índice de Covid-19, agrupam as Classes Econômicas de número 6 e 0, que contemplam atividades de Comércio, Atividades Essenciais e Administração Pública. Também observa-se que a Classe Econômica de número 2, que abrange, principalmente, o setor Agropecuário, teve bastante isolamento social, se comparado a outras classes econômicas.

Analisando o gráfico e as hipóteses levantadas neste trabalho nota-se tendência negativa para a Hipótese 1, visto que as cidades com perfil econômico predominantemente do setor secundário que estão contemplados nas classes 1, 3 e 4, não possuem altos índices de casos de Covid-19, mesmo com uma taxa de isolamento social dissipada.

Para a Hipótese 2, existe tendência afirmativa, tendo em vista que as classes que contemplam o setor terciário, obtiveram altos índices de casos de Covid-19. A Hipótese 3, também tem tendência afirmativa, visto que a classe do setor primário / agricultura (Classe 1) tem baixo índice de casos de Covid-19, e isso é atrelado a altas taxas de isolamento social.

CONSIDERAÇÕES FINAIS

Técnicas de *machine learning* tem grande importância na área de tecnologia da informação e seu uso é cada vez mais comum no desenvolvimento de *software*. Através de modelos, a inteligência artificial facilita o encontro de respostas para problemas que são dificilmente interpretados por seres humanos devido a dificuldade de se analisar grandes cargas de dados.

Este trabalho implementa um modelo de *machine learning* a fim de correlacionar dados de isolamento social, casos de Covid-19 e classe econômica dos municípios do Estado do Rio Grande do Sul. Por meio da interpretação dessa correlação, foi possível responder algumas hipóteses levantadas.

Notou-se que cidades com perfil econômico predominantemente rural têm menos casos de Covid-19 e mais isolamento social. Cidades com o perfil econômico do setor que corresponde a indústria apresentaram distanciamento social e casos de covid-19 dissipados. Cidades com perfil econômico no setor de serviços, apresentaram tendência de mais casos de Covid-19.

AGRADECIMENTOS

Este trabalho é financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes).

REFERÊNCIAS BIBLIOGRÁFICAS

MOURA, R.; FERRAZ, L. Meios de Controle à Pandemia da COVID-19 e a Inviolabilidade da Privacidade. InLoco Report. 2020.

REHBEIN, Matheus H.. Comparação de Métodos Não Supervisionados: Um Caso Baseado no CNAE 2.0 e na Covid-19. 2020. Relatório Técnico. Programa de Pós Graduação em Modelagem Matemática e Computacional, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2020.