

UM ESTUDO DE CLUSTERIZAÇÃO DE MICRO E MACRORREGIÕES COM BASE EM INDICADORES DE CONTAMINAÇÃO DAS POPULAÇÕES PELO CORONAVÍRUS E DE SUAS CORRELAÇÕES COM VARIÁVEIS SOCIOECONÔMICAS E GEOGRÁFICAS

Daniele Alves Deutner e Prof. Dr. Arnaldo Rabello de Aguiar Vallim Filho

Apoio: PIVIC Mackenzie

RESUMO

Nesta pesquisa, foi realizado o desenvolvimento de um conjunto de clusters de regiões que possuem comportamento semelhante em relação a variáveis socioeconômicas e paralelamente construíram-se clusters agrupando regiões que caracterizam o desenvolvimento similar da doença COVID-19. Os dados utilizados representam o período de isolamento de um ano, março de 2020 a março de 2021 e foram levantados para uma base das 27 Unidades da Federação do Brasil e também para todos os países do mundo. Os dois tipos de construções de clusters foram comparados, mostrando similaridade, seja para os estados do Brasil, seja para países do mundo. Técnicas adicionais foram aplicadas envolvendo análise de correlação, árvore de regressão e equações de regressão, e esses resultados foram confirmados. Entende-se que estes resultados podem ser utilizados para o desenvolvimento de modelos de previsão que podem levar à definição de políticas públicas para diferentes regiões.

Palavras-chave: Clusterização, Análise de Correlação, Regressão, COVID-19

ABSTRACT

ABSTRACT

In this research, a set of clusters of regions that have similar behavior in relation to socioeconomic variables were developed and, in parallel, clusters were built grouping regions that characterize the similar development of the COVID-19 disease. The data used represent the isolation period of one year, March 2020 to March 2021 and were collected for a base of the 27 Federation Units of Brazil and also for all countries in the world. The two types of cluster construction were compared, showing similarity, whether for the states of Brazil or for countries in the world. Additional techniques were applied involving correlation analysis, regression tree and regression equations, and these results were confirmed. It is understood that these results can be used to develop forecast models that can lead to the definition of public policies for different regions.

Keywords: Clustering, Correlation Analysis, Regression, COVID-19

1. INTRODUÇÃO

A nova doença que afetou todo o mundo foi a Covid-19, que teve seus primeiros aparecimentos na China em 2019 e foi se espalhando pelo mundo inteiro chegando ao Brasil em março de 2020. Para evitar mais a contaminação, foi decretado o isolamento social a modo de tentar diminuir a transmissão do vírus em todos os lugares.

Uma questão fundamental, que se coloca é o estudo das diferentes bases de dados existentes, e são muitas em todo o mundo, que apresentam uma diversidade de variáveis, de forma a identificar aquelas que seriam mais relevantes na caracterização de clusters, em que os elementos desses clusters tivessem comportamento similar, em relação a essas variáveis.

Note-se que com o aparecimento desse vírus respiratório, o mundo inteiro praticamente parou por conta da pandemia, com prejuízos ainda incalculáveis sob vários pontos de vista. Assim, quanto mais conhecimento se tiver sobre a pandemia e suas características, mais eficientes e eficazes poderão ser as ações de planejamento e combate à doença.

Nesse sentido, a Ciência de Dados pode ser uma grande aliada. Com as técnicas disponíveis hoje, consegue-se analisar com eficiência e eficácia os dados obtidos, e em particular, com o auxílio de técnicas de clusterização, pode-se agrupar dados, e agilizar as análises, separando-as em grupos semelhantes.

Portanto, conseguindo-se desenvolver agrupamentos de regiões similares em relação à doença, pode-se estabelecer ações de planejamento para esta pandemia e, ainda, possivelmente, utilizar as mesmas técnicas e metodologia para pandemias futuras que poderão ocorrer.

O que se busca, portanto, é o uso da Ciência de Dados para agrupar regiões que sejam similares em relação à doença e em relação a aspectos socioeconômicos.

Assim, é possível formular a seguinte pergunta de pesquisa, que representaria bem o que se pretende neste estudo

Seria possível identificar um grupo de variáveis, relativas a regiões estabelecidas, que caracterizam aspectos da pandemia de COVID-19, de forma a se estabelecer grupos de regiões com comportamentos similares, seja em características socioeconômicas, seja em relação à COVID-19, caracterizando, assim, cada um desses grupos como clusters, do ponto de vista de Ciência de Dados?

Mais um novo vírus que se espalhou pelo mundo inteiro, está tomando conta das últimas notícias de 2020. Todos os dias, saem comunicados sobre os *status* de evolução do

vírus e no momento que se finaliza este texto, já foi ultrapassada a marca de 900 mil pessoas infectadas pelo coronavírus no Brasil (MINISTÉRIO DA SAÚDE, 2020).

Este é um problema gravíssimo do ponto de vista de saúde pública, assim como, do ponto de vista econômico, com a redução drástica das atividades econômicas em todo o planeta.

Assim, com a ajuda da Ciência de Dados, um estudo que possibilite analisar as grandes e inúmeras bases de dados diariamente atualizadas, pode-se ter uma visão bem mais avançada sobre essa pandemia.

Uma das ferramentas da Ciência de Dados é a clusterização dos dados e com ela ajudará a mostrar as micro e macrorregiões com base em indicadores de contaminação das populações pelo coronavírus e mostrará quais são as regiões mais e menos afetadas pela pandemia. Uma correlação dessas regiões com outras variáveis poderá ser de grande valia para a definição de ações específicas para cada categoria de regiões, diferentemente de hoje, que se vê praticamente um único leque de ações, fortemente baseadas em distanciamento social, higiene e uso de máscaras. Identificando-se estratos das populações, que tenham resposta similar à doença, e que possuam características similares, poderá gerar ações direcionadas especificamente para determinados segmentos da população que poderão ter mais eficácia.

Do ponto de vista científico, este projeto levará ao aprofundamento dos estudos sobre técnicas e algoritmos de clusterização, e de suas ferramentas computacionais, o que trará grande ganho de conhecimento e maior domínio dessa tecnologia, que pode ser aplicada em praticamente qualquer área de atividade, como poderá ser visto na seção 2. A Ciência de Dados é hoje definitivamente uma área em grande ascensão, que se desenvolve e se aprimora rapidamente e se espalha com aplicações por toda a sociedade. O aprofundamento de estudos e pesquisas nessa área, permitindo forte domínio de toda essa tecnologia, é, assim, fundamental para uma universidade.

Entende-se, portanto, que este é um projeto que pode trazer contribuições relevantes sob diferentes aspectos e é assim, plenamente justificado.

O objetivo geral deste projeto foi o desenvolvimento de um conjunto de clusters de regiões que possuem comportamento semelhante em relação a variáveis que caracterizam o desenvolvimento da doença COVID-19.

Uma questão fundamental, seria estudar essas variáveis, de forma a identificar aquelas que seriam mais relevantes na caracterização de clusters em que os seus elementos tivessem comportamento similar, em relação a essas variáveis.

Estudar as possíveis correlações desses clusters com características socioeconômicas e geográficas das regiões.

Desenvolver o estudo em termos de microanálises, envolvendo municípios e comunidades, à medida que se tenha disponibilidade de dados, e também, em termos de macro análises, envolvendo estados e países. Assim, em termos de Objetivos Específicos, pretende-se:

- Experimentar com diferentes técnicas de clusterização em bases de dados públicas relativas à pandemia;
- Desenvolver experimentos com bases de dados associadas a macrorregiões, envolvendo países;
- Desenvolver experimentos com bases de dados associadas a macrorregiões, envolvendo estados brasileiros;
- Desenvolver experimentos com bases de dados associadas a microrregiões, envolvendo pequenos municípios brasileiros e distritos de grandes cidades;
- Identificar correlações dos clusters definidos, com variáveis socioeconômicas.
- Comparar os resultados por meio de visualizações e análises estatísticas.

2. REFERENCIAL TEÓRICO

Este capítulo tem como objetivo apresentar uma revisão da literatura de trabalhos científicos que possuem relação com o tema deste projeto. As obras estudadas auxiliam no entendimento da problemática e no conhecimento das técnicas associadas aos objetivos da pesquisa. Neste caso, tem-se duas áreas relevantes de pesquisa: o próprio contexto da pandemia do covid-19 e o estudo das técnicas de aprendizado de máquina relativas a clusterização.

2.1 Pandemia de COVID-19

A pandemia de COVID-19, foi causada pelo novo coronavírus, que é um vírus que causa infecções respiratórias, e que tiveram diagnosticados os primeiros casos na China no final de 2019. Conforme definição do Ministério da Saúde (2020), a COVID-19 é uma doença causada pelo coronavírus SARS-CoV-2, que apresenta um quadro clínico que varia de infecções assintomáticas a quadros respiratórios graves. De acordo com a Organização Mundial de Saúde - OMS (*World Healthy Organization*, <https://www.who.int/>), a maioria dos pacientes com COVID-19 (cerca de 80%) podem ser assintomáticos e cerca de 20% dos casos podem requerer atendimento hospitalar por apresentarem dificuldade respiratória e desses casos aproximadamente 5% podem necessitar de suporte para o tratamento de insuficiência respiratória (suporte ventilatório)".

Há falhas nas estatísticas brasileiras em relação à pandemia, como por exemplo na parte de testagem, pois estão realizando os testes em pacientes que são considerados grupo de risco e não praticando nos pacientes em geral. O índice no país é muito baixo, conforme pode-se verificar na figura 1.



Figura 1 – Índice de Testes COVID-19 por 100 mil habitantes¹

O vírus se espalhou rapidamente em todo o país em um curto período, com concentração no sudeste, mas também em estados do norte e nordeste (PEDROSA, R. 2020). A figura 2 mostra a evolução no país e no mundo, onde pode-se observar que no Brasil a doença está em ascendência enquanto outros países já tiveram queda, algumas acentuadas.

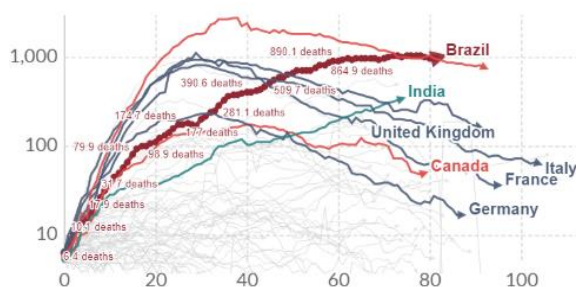


Figura 2 – Evolução² de Casos de COVID-19 no Mundo

No momento em que este texto foi finalizado, os dados mostram que o país tem mais de 850 mil casos confirmados da COVID-19, dos quais mais de 43.000 são óbitos. Os números nacionais apontam para uma taxa de mortalidade de 4,999%. (MINISTÉRIO DA SAÚDE, 2020).

Há muitos estudos sobre o tema, como o de Albuquerque et al. (2020) que mostra que no Brasil, diversas medidas vêm sendo implementadas, nas três esferas de governo, variando de medidas de estímulo ao isolamento social à construção de infraestruturas temporárias de atendimento – como os hospitais de campanha – para lidar com o aumento da busca pelos

¹ <https://ourworldindata.org/coronavirus>

² <https://github.com/owid/covid-19-data>

serviços de saúde. Alguns estudos mostram que se o Brasil tivesse adotado o distanciamento social mais cedo, a rotina normal retornaria em poucos meses conforme o levantamento por (BENARRÓS, 2020). Um debate sobre os efeitos econômicos e sociais da pandemia e de seu enfrentamento tem sido travado, em particular, sobre a questão social (PAIVA et al., 2020). O pagamento de valor à vulneráveis, uma medida é um desses pontos (ALBUQUERQUE et al., 2020).

A questão territorial da pandemia, sobretudo no que diz respeito aos espaços intermunicipais das principais metrópoles brasileiras é outro aspecto importante que vem sendo estudado, na medida da relação existente entre a taxa de contaminação, a disponibilidade da infraestrutura urbana e a densidade demográfica (CONNOLLY et al., 2020).

No Brasil, estima-se que metade dos assentamentos urbanos seriam precários, irregulares ou informais, o que agrava ainda mais o quadro. Mesmo sem estimativas precisas sobre a ocorrência de assentamentos precários e informais, sabe-se, pela experiência vivida na epidemia de SARS, em 2003, que as condições socioeconômicas, de habitação e de acesso à infraestrutura precárias contribuem para ampliar a vulnerabilidade de contaminação, exigindo medidas específicas para as diferentes clusters do território (CABALLERO-ANTHONY, 2003), que é o foco deste projeto.

Há, portanto, quantidade significativa de dados disponíveis e de literatura sobre o tema, o que é um fator que deverá favorecer o desenvolvimento deste projeto.

2.2 Clusterização

Clusterização é um processo de agrupamento dos algoritmos / dados em grupos semelhantes. O estudo de clusterização é antigo. A Estatística é quem primeiro estudou o problema (EATON, 1983), porém mais recentemente com o crescimento das aplicações de computação o assunto ganhou novo impulso, em particular, com o crescimento da Ciência de Dados.

Pode-se dizer que “clusterização” é o processo de organizar elementos em grupos cujos membros são similares de alguma forma (MATTEUCCI, 2006).

Segundo Ochi et al. (2005), o Problema de Clusterização (PC) consiste em dada uma base de dados X , agrupar (clusterizar) os objetos (elementos) de X de modo que objetos mais similares entre si sejam alocados ao mesmo cluster, enquanto, objetos menos similares sejam alocados em outros clusters.

Em termos de abordagens, pode-se ter o problema de K-Clusterização com K pré-definido. Já quando K não é previamente conhecido, tem-se o Problema de Clusterização Automática (PCA). O que caracteriza este tipo de problema em mineração de dados, é que

nesta área a base de dados costuma ter um porte elevado e o número de atributos de cada elemento é, em geral, alto (OCHI et al., 2005).

Há dois tipos principais de métodos: o hierárquico e o de particionamento. No método do tipo Particionamento, parte-se de um conjunto “K” de clusters e procede-se a um intercâmbio de elementos entre esses clusters. O tipo hierárquico pode ser de Aglomeração ou de Divisão.

Em Aglomeração (*bottom-up*) pode-se iniciar com cada elemento se constituindo em um cluster. A seguir, os clusters vão sendo unidos. A união de clusters sempre ocorre com base em alguma medida de similaridade. No processo de Divisão (*top-down*), se inicia com um só cluster e procede-se a uma sequência de divisões, com base em algum critério (SILVA et al., 2016; CASTRO et al., 2016).

Há quatro algoritmos de clusterização que se destacam: K-Médias, Fuzzy C-means, a Clusterização Mista Gaussiana e a Clusterização Hierárquica, cujo processo já foi explicado acima. Sobre os demais, o K-Média (K-means) é o mais conhecido e é um algoritmo de particionamento, conforme já colocado acima. Neste algoritmo, parte-se de um conjunto “K” de centróides e vai-se agregando elementos a cada centróide. Posteriormente, procede-se a um intercâmbio de elementos entre os clusters, de forma se tentar melhorar uma medida de similaridade, que foi definida como critério de agrupamento. Em geral, esta medida é a distância entre os elementos (ZHAO, 2015). Já o Fuzzy C-means é um algoritmo de agrupamento sobreposto, e a Mistura de Gaussiana é um algoritmo probabilístico de agrupamento”. (MATTEUCCI, 2006).

Os métodos de particionamento são muito usados, como as técnicas *k-medoids* e *k-médias*. A diferença entre elas é que no *k-medoids* o elemento que representa o cluster, é definido de acordo com seus atributos sem que haja muita influência dos valores limites do cluster. Já no caso do *k-médias* o centróide dos seus elementos representa o cluster, e corresponde ao valor médio de cada atributo considerado (OCHI et al., 2005).

Para conseguir um bom cluster, precisam que os objetivos sejam detalhados corretamente, com isso o usuário informará seu critério e o retorno será atendido com a sua expectativa, por exemplo, “poderíamos estar interessados em encontrar representantes para grupos homogêneos (redução de dados), em encontrar “aglomerados naturais” e descrever suas propriedades desconhecidas (tipos de dados “naturais”), em encontrar agrupamentos úteis e adequados (classes de dados “úteis”) ou na localização de objetos de dados incomuns (detecção de outlier)”. (MATTEUCCI, 2006).

Segundo Matteucci (2006), há diversos problemas na geração de clusters entre eles se destaca o fato de se ter de lidar com altas dimensões (muitos atributos dos dados) e grande número de observações (exemplares). Isto pode ser problemático devido à questão do tempo e do resultado do algoritmo de agrupamento, que pode ser interpretado de diferentes formas.

Um aspecto importante é a questão dos critérios de definição dos clusters mais próximos entre si, principalmente na clusterização hierárquica. Nesse sentido, há três tipos de ligações: simples ou única, a completa e a média.

Conforme Johnson, 1967), que foi o criador deste método de clusterização em 1967, na ligação-simples (*single-linkage clustering*) a distância entre clusters é a menor distância entre membros de um cluster e do outro, também chamado de conexão ou método do mínimo. O autor afirma, que o critério de ligação única tende a produzir clusters irregulares, não muito homogêneos internamente. Já na ligação completa (*complete-linkage clustering*) a distância entre clusters é a maior distância entre qualquer membro de um cluster e do outro, e é também chamado de diâmetro ou método do máximo. Na ligação média (*average-linkage clustering*) a distância entre clusters é a distância média entre os membros de um cluster e do outro (JOHNSON, 1967).

Um aspecto que deve ser considerado na questão de clusterização é que esta é uma técnica de Aprendizado de Máquina (*machine learning*), que segundo Chollet (2018) é uma forma de Inteligência Artificial (IA) que permite que um sistema aprenda com os dados, e não através de programação explícita, como é o caso do ramo da IA simbólica. Tem-se dois tipos de aprendizado: supervisionado e não supervisionado. O objetivo do aprendizado supervisionado (AS) é encontrar padrões nos dados, que posteriormente, podem ser aplicados a outros dados. O AS parte de um conjunto de dados em que já exista uma classificação das observações, e com isto, calibra (“treina”) o algoritmo. A clusterização se enquadra no aprendizado não supervisionado (AnS), cujo objetivo é encontrar relações de similaridade entre os dados, que é justamente o caso da clusterização. Ao contrário do anterior, no AnS não há qualquer tipo de classificação ou organização dos dados (SILVA et al., 2016; CASTRO et al., 2016).

Estas técnicas de clusterização podem ser utilizadas no processo de decisões importantes, associadas a custos muito elevados. Khalid e Herbert Hansen (2018), por exemplo, utilizaram aprendizado de máquina não supervisionado para clusterizar instalações internacionais de uma organização, e com isso, apoiar a decisão de localização de novas instalações. No estudo, um total de 24 indicadores de 94 países, categorizados em cinco grupos, foram utilizados na análise. Após o agrupamento, estes foram comparados e registrados para selecionar os países. Os autores afirmam que a técnica foi bem efetiva

fornecendo uma saída quantitativa sem subjetividade apoiando os tomadores de decisão sobre os locais das instalações.

Já Resta et al. (2017) utilizaram a técnica para clusterização de fluxos de pacientes em emergências. O estudo mostra que com grande movimentação de internações, os hospitais têm uma demanda elevada no setor de emergência, com pacientes que aguardam para serem admitidos nas enfermarias dos hospitais. Os autores afirmam que o estudo se justifica pelas restrições orçamentárias públicas que implicam em reduzir o número de leitos de internação. Para melhorar o planejamento das ofertas de leitos, deve-se atentar para os gargalos do sistema, como o internamento no pronto-socorro de pacientes emergência. Neste caso, utilizaram a técnica redes neurais artificiais (RNA) com os mapas auto-organizáveis (SOM – *self organizing maps*).

Esse é um tipo de RNA, proposta por KOHONEN (1982) é um tipo de rede neural artificial em formato de matriz onde cada neurônio forma um agrupamento com os dados mais próximos entre si. Nesse sentido, o SOM é um gráfico de similaridade onde cada neurônio está posicionado em uma coordenada definida pelo centróide dos dados do seu cluster. Resta et. al. (2017) afirmam que o SOM foi usado extensivamente para classificar dados por meio de representação bidimensional e criar clusters capazes de agrupar características comuns dos dados de entrada em subconjuntos.

Uma outra aplicação levantada na literatura se refere à área de energia. Liu et. al. (2018) utilizaram a técnica para clusterizar setores industriais na China, e com isso, medir a quantidade de energia utilizada nas empresas. Isto foi feito com base no K-means, com regras de associação multidimensional, para analisar as características do consumo regional de energia em diferentes indústrias e empresas. Segundo Liu et. al. (2018), o algoritmo agrupa indústrias e empresas regionais com alto ou baixo consumo de energia.

2.3. Análise de Correlação

Análise de Correlação é um valor (X) de uma variável que quando ela aumenta/diminui, o outro valor (Y) tende a aumentar/diminuir também (MIOT, 2018). Para ter uma boa análise de correlação entre duas variáveis quantitativas (X e Y) deve ser utilizado o diagrama de dispersão, que auxilia a identificar se tem relação dessas duas variáveis para o mesmo conjunto de dados.

Com os resultados que diz se tem ou não tem correlação, o seu coeficiente de correlação vai de -1 a +1. Quanto mais próximo do valor do coeficiente for do 1, mais forte é a relação entre as duas variáveis, tanto positivas como negativas.

2.4. Árvore de Regressão - Regression Tree

A árvore de regressão é um exemplo de árvore de decisão que emprega a soma dos quadrados e ela também prevê os valores dos campos alvos e com isso, os resultados dessas combinações serão confirmados pelos valores das entradas.

Uma árvore de regressão calcula um valor médio previsto para cada nó da árvore. Cada nó é quebrado em dois ou mais nós filhos para limitar a soma dos quadrados para o nó. Os nós inferiores equivalem a um tipo de preditores que são ligados quando existe o aumento da soma dos quadrados que é razoável dentro do limite determinado (ARGENTON, 2013)

Para começar a construir uma árvore de decisão, precisasse de um nó pai (raiz) e para cada nó é dividido em dois nós filhos (folhas) e assim por diante até que não consiga mais ser dividido.

2.5. Regressão Linear - Função GLM

Regressão Linear tem como objetivo verificar a relação de uma variável resposta e uma ou mais variáveis explicativas, para identificar uma função que a descreva (OLIVEIRA, 2019). Por meio dela é capaz de entender o porquê existem variação de um determinado fenômeno e fazer uma previsão do seu comportamento conforme as variáveis explicativas. Se existe somente uma variável explicativa ela é considerada uma regressão simples, por outro lado, se existir várias variáveis explicativas ela é considerada uma regressão múltipla.

O conceito de Modelo Lineares Generalizados (GLM) é uma ampliação dos modelos de regressão simples e múltiplas. Com essa função, consegue auxiliar em outras distribuições para verificar erros ou variáveis que se conectam na média das variáveis resposta com as probabilidades lineares das variáveis explicativas.

Segundo Cordeiro, existem casos especiais do GLM que são eles:

- modelo clássico de regressão múltipla (Legendre, Gauss, início do século XIX) e modelo de análise de variância para experimentos planejados (Fisher, 1920 a 1935) com o erro aleatório tendo distribuição normal;
- os modelos de testes de vida, envolvendo a distribuição exponencial (Feigl e Zelen, 1965; Zippin e Armitage, 1966; Gasser, 1967);
- Modelo de Análise de variância com efeitos aleatórios;
- Modelo Estrutural para Dados com Distribuição Gama;
- Modelo de Regressão não-simétrico.

3. METODOLOGIA

A pesquisa foi desenvolvida é um tipo de pesquisa aplicada, já que sua utilização prática é total. O estudo adotou uma abordagem quantitativa, e sua finalidade é o aperfeiçoamento e ampliação de uma ferramenta computacional, incluindo a implementação dessas melhorias. Como finalidade tem-se ainda, a obtenção de conhecimento novo sobre o comportamento de diferentes regiões geográficas em relação à pandemia de Coronavírus.

Em termos de etapas do estudo, teve-se:

- . Pesquisa Bibliográfica
- . Levantamento de Dados
- . Pré Processamento de Dados
- . Desenvolvimento de Modelos e Experimentos
- . Análise de Resultados
- . Documentação

4. RESULTADO E DISCUSSÃO

Esta seção apresenta os experimentos desenvolvidos no estudo, as tecnologias utilizadas, os resultados obtidos e uma discussão sobre esses resultados

4.1. Bases de Dados e Pré processamento

Para o estudo foram levantados dados relativos à COVID 19 e dados socioeconômicos de diferentes regiões do país e do mundo. Na questão interna do país, as informações dizem respeito aos estados do Brasil (Unidades da Federação UFs), e no caso do mundo os dados se referem aos países.

Para esses levantamentos foram utilizadas diversas plataformas:

SEADE – Fundação Sistema Estadual de Análise de Dados : <https://www.seade.gov.br/>

IBGE – Instituto Brasileiro de Geografia e Estatística: <https://www.ibge.gov.br/>

Repositorio Datasharing Fapesp: <https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/2>

Com dados de: . Hospital Israelita Albert Einstein

. Grupo Fleury

. Hospital Sírio-Libanês

OMS – Organização Mundial da Saúde: <https://www.who.int/>

Ministério da Saúde: <https://covid.saude.gov.br/>

Para os dois casos, países e UFs do Brasil, foram levantados os mesmos dados socioeconômicos, para o ano de 2020, os quais são apresentados abaixo:

Nome do País ou UF

Nome da Capital do País ou UF

Latitude da Capital do País ou UF

Área (Km2)

Demografia (hab/km2)

População

PIB

Renda per capita

Índice de Desemprego

Índice Gini

Índice de Mortalidade Infantil

Idade Média da População

Expectativa de Vida da População

Índice de Analfabetismo

Anos de Escolaridade da População

No caso de países foi ainda levantado o Continente do País.

Os dados foram todos obtidos nas diferentes fontes no formato .csv ou .xlsx, ambos em planilhas Excel, e antes do uso desses dados para o desenvolvimento das análises estes passaram por alguns tratamentos para limpeza de ruídos e melhoria da qualidade geral.

- Selecionou-se nas planilhas originais os dados a serem convertidos
- Com os dados selecionados, utilizou-se um dos atalhos do teclado pouco conhecido, mas que traz um ganho excepcional, ALT pressionado mais a digitação da palavra DEC que realizou a conversão instantaneamente.

Nos casos das cedilhas e acentos que tiveram que ser removidos no arquivo Excel, foi preciso criar uma função, que remove acentos. A criação de funções é feita pressionando-se ALT + F11, o que faz surgir uma nova janela do editor Visual Basic.

No menu acesse INSERIR/MÓDULO, foi construída a função descrita abaixo:

CAcento = "àáâãäåèéêëîïñòóôõöùúüÄÅÃÄÈÉÊËÏÎÏÒÓÔÕÖÙÚÛÜÇçÑñ"

SAceto = "aaaaaeeeeiiiiiooooouuuuuAAAAEEEEIIIOOOOUUUUcCnN"

```
Texto = ""
```

```
if Palavra <> "" then
```

```
For X = 1 to Len(Palavra)
```

```

Letra = mid(Palavra,X,1)
Pos_Acento = inStr(CAcento,Letra)

```

```

if Pos_Acento > 0 then

```

```
Letra = mid(SAcento,Pos  Acento,1)
```

end if

Texto = Texto & Letra

next

TiraAcento = Texto

end if

end function

Function VerificaPalavra(atributo)

```

Dim i
  Dim id
  Dim Auxiliar
  Dim Resultado
  Auxiliar = Split(Atributo, " ", - 1, vbBinaryCompare)
  For i = LBound(Auxiliar) To Ubound(Auxiliar)
    Resultado = Resultado & " " & TiraAcento(Auxiliar(i))
  Next
  VerificaPalavra = Trim(Resultado)
end function

```

Feito isso, bastou usar a função TiraAcento(), na própria janela do Excel (ADMIN, 2021).

4.2 Análise e Comparação de Clusters

Foram desenvolvidos clusters baseados em variáveis socioeconômicas e em variáveis COVID (No. de casos e No. de óbitos), tanto para os estados do Brasil como para os países do mundo. Para isto, foram utilizadas bibliotecas da linguagem R. No caso foi utilizado o algoritmo de K-médias (OCHI et al., 2005), discutido na seção 2. Com os clusters desenvolvidos para diferentes combinações de variáveis socioeconômicas, os resultados foram comparados com clusters baseados nas variáveis COVID. Os resultados mostraram uma similaridade considerável em boa parte dos casos, ou seja, clusters construídos baseados em variáveis socioeconômicas eram similares a clusters construídos com base nas variáveis COVID.

Os gráficos a seguir procuram retratar essa comparação.

4.2.1. Brasil

Cada ponto representa uma UF associada a um cluster e apresenta-se seis tipos de clusters: baseados em coordenadas geográficas, em área (km²), em população, em renda, em expectativa de vida e em escolaridade. Os gráficos de cada tipo de cluster são apresentados ao lado dos clusters gerados com base nas variáveis COVID, para comparação.

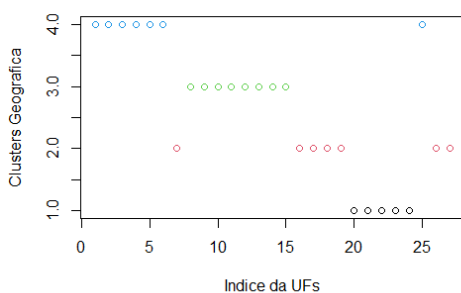


Figura 4.1.1a: Clusters com base Geográfica

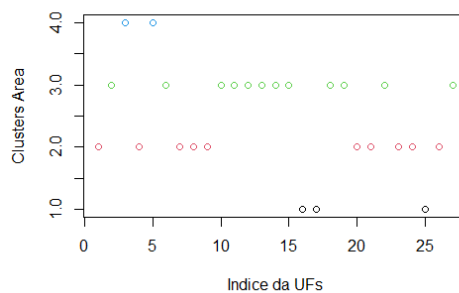


Figura 4.1.2a: Cluster baseados em Area

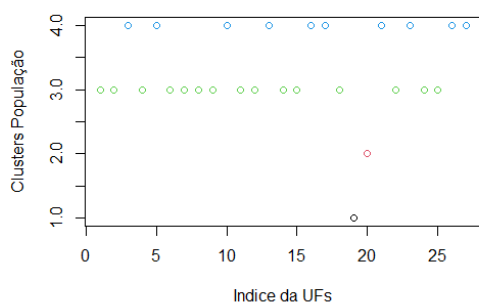


Figura 4.1.3a: Cluster População

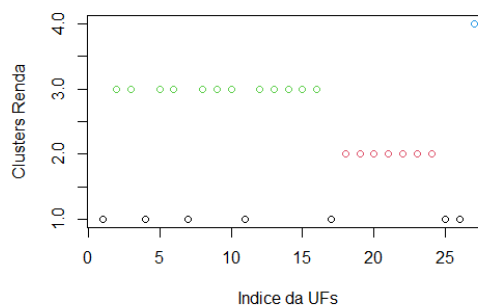


Figura 4.1.4a: Cluster Renda

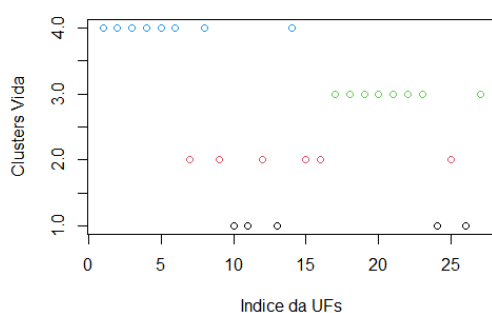


Figura 4.1.5a: Cluster Vida

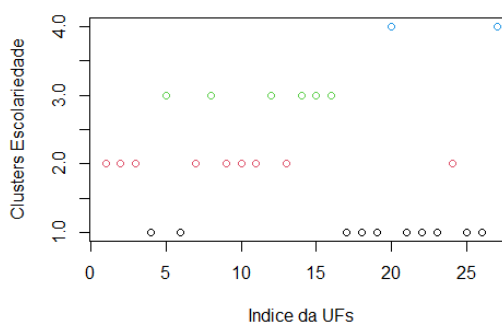


Figura 4.1.6a: Cluster Escolaridade

Os resultados sintéticos dessas análises podem ser expressos em termos de um índice de coincidências de pontos em clusters “socioeconômicos” e clusters “COVID”. O índice verifica o percentual de pontos que se encontram juntos no mesmo cluster nos dois tipos de construção. Os resultados obtidos foram: Cluster Geográfico: 67%, Cluster Área: 67%, Cluster População :89%, Cluster Renda: 59%, Cluster Vida: 48% e Cluster Escolaridade: 70%. Todos comparados com os clusters COVID.

4.2.2. Mundo

No caso do mundo foi feito o mesmo tipo de comparação, com os gráficos de cada tipo de cluster baseado nas variáveis socioeconômicas sendo apresentado ao lado dos clusters gerados com base nas variáveis COVID.

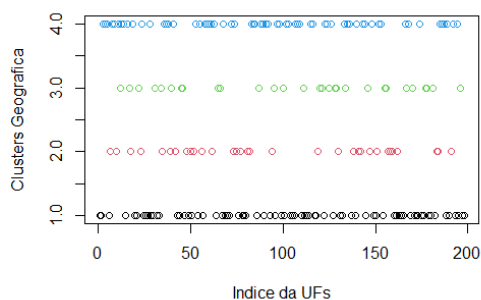


Figura 4.1.1a: Cluster Geográfico

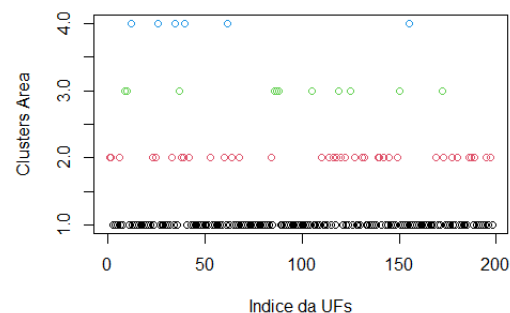


Figura 4.1.2a: Cluster Area

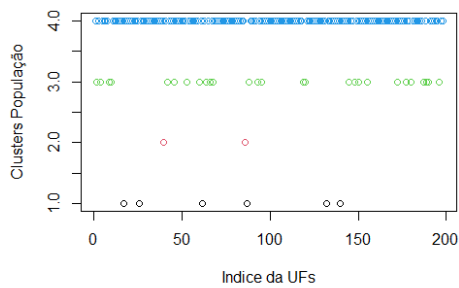


Figura 4.1.3a: Cluster População

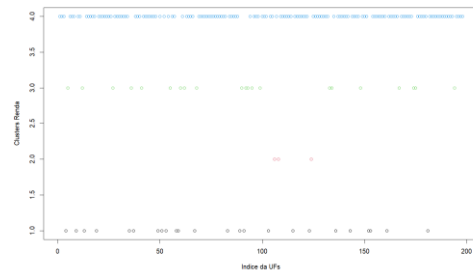


Figura 4.1.4a: Cluster Renda

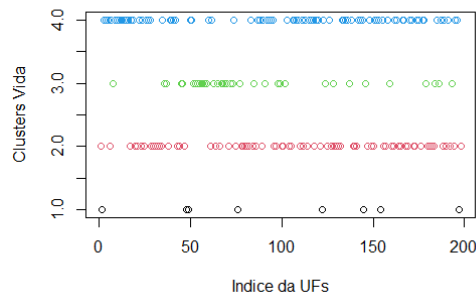


Figura 4.1.5a: Cluster Vida

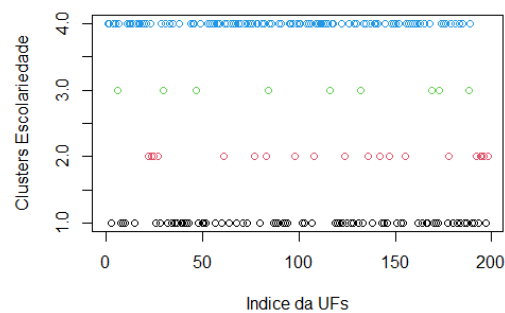


Figura 4.1.6a: Cluster Escolaridade

Neste caso, as comparações dos dois tipos de construções de clusters mostraram os seguintes índices de coincidência: Cluster Geográfico: 70%, Cluster Área: 91%, Cluster População: 95%, Cluster Renda: 94%, Cluster Vida: 90% e Cluster Escolaridade: 82%. Os índices mostram, portanto, uma similaridade ainda maior do que no caso das UFs do Brasil

4.3 Análises Adicionais

Além da clusterização foram desenvolvidas análise de correlação entre as variáveis COVID com as demais, árvore de regressão para prever as variáveis COVID e equações de regressão linear (modelo gaussiano) e equação de regressão com hipótese de comportamento de distribuição de Poisson para os dados. Neste caso, fez-se uso do modelo GLM – Generalized Linear Model, para o qual há bibliotecas disponíveis na linguagem R. Os resultados são apresentados nas próximas seções.

4.3.1 Análises adicionais - UFs Brasil

Tanto no UF Brasil quanto no Mundo, citar que nas figuras 4.2a, 4.2b e 4.2c (análise de correlação) os pontos azuis em diagonal não devem ser levados em consideração, pois eles são resultados comparativos com eles mesmos, por isso devem ser ignorados. Os pontos que foram agrupadas conformes as cores, a figura 4.2a é uma análise completa, a figura 4.2b a análise simplificada da figura 4.2a e a figura 4.2c seria a análise das variáveis mais impactadas pela covid de número de casos total e de mortes, nas figuras 4.2d e nas figuras

4.2e são diagramas para auxiliar na clusterização dos dados de número de casos total e de mortes com os restos dos dados, tinham uma população densa e grande, como por exemplo o estado de São Paulo. Em contrapartida os estados que tiveram o menor casos e óbitos de covid tem-se um baixo índice de analfabetismo entre eles como no caso do estado de Santa Catarina que a mesma relação se encontra quando se analisa os dados de índice de desemprego, como ocorre no estado do Rio Grande do Sul.

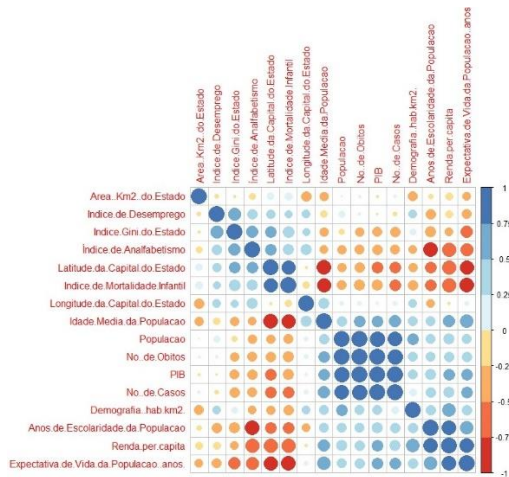


Figura 4.2a: Saídas Corr Full

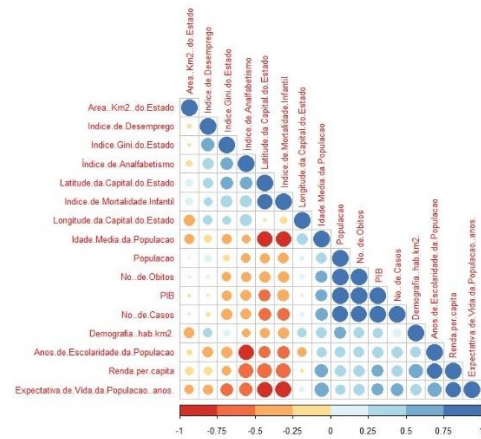


Figura 4.2b: Saídas Corr Lower



Figura 4.2c: Saídas Corr Variáveis Covid

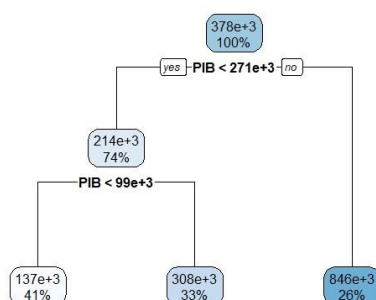


Figura 4.2d: Saídas RT Casos

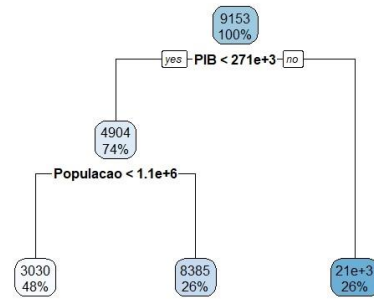


Figura 4.2e: Saídas RT Óbitos

4.3 UF Mundo

Na figura 4.3a tem duas árvores com modelos de árvore de regressão, e neles se confirma que os países que mais tiveram casos e mortes por covid-19, tinha uma população densa e grande, como por exemplo os Estados Unidos. Em contrapartida os estados que tiveram o menor casos e óbitos de covid tem-se um baixo Índice de Gini do país entre eles como no caso na Eslovênia que a mesma relação se encontra quando se analisa os dados de índice de analfabetismo, como ocorre na Noruega.

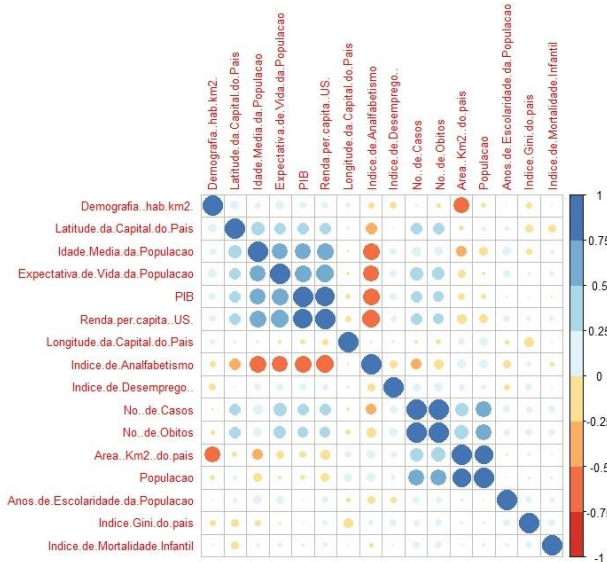


Figura 4.3a: Corr Full

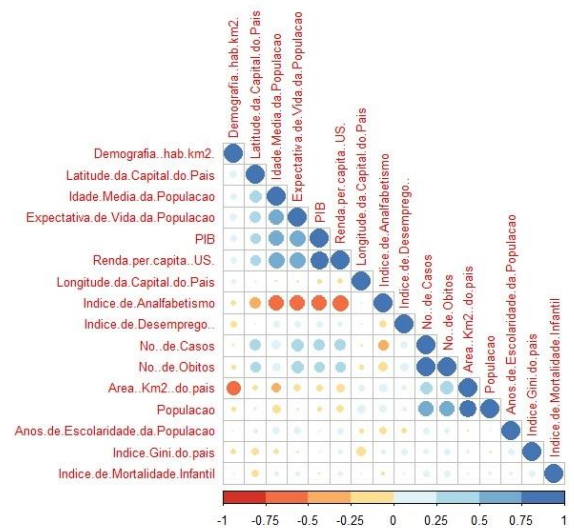


Figura 4.3b: Corr Lower



Figura 4.3c: Corr Variáveis Covid

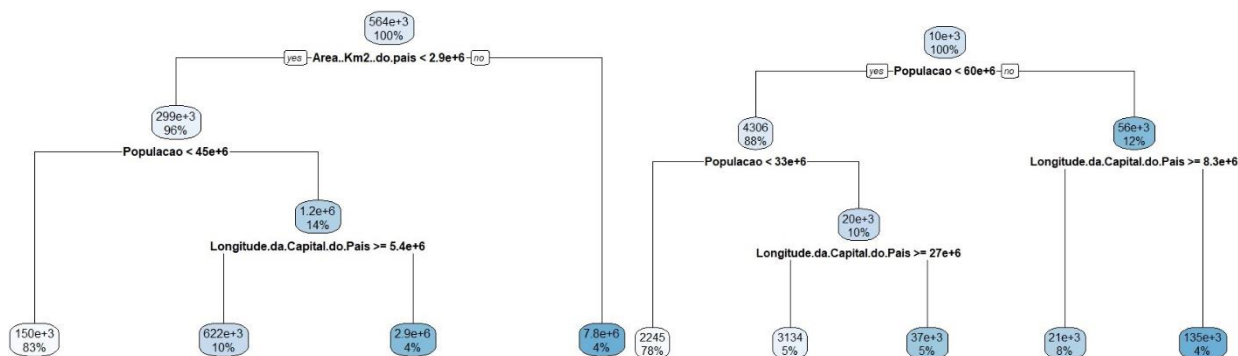


Figura 4.3.d: Árvore de Regressão - Casos

Figura 4.3.e: Árvore de Regressão - Óbitos

5. CONSIDERAÇÕES FINAIS

O estudo demandou um longo levantamento de dados e um trabalho de limpeza e transformação para melhoria de qualidade dessas informações. A partir dessa base, um conjunto de técnicas foram aplicadas aos dados, inicialmente por meio de clusterização, onde foi possível verificar a similaridade entre clusters gerados com base em variáveis associadas diretamente `COVID e outros clusters construídos com base em variáveis socioeconômicas. Posteriormente, análise adicionais foram desenvolvidas, correlação, árvore de regressão e equações de regressão, e todas confirmaram os resultados obtidos pela clusterização.

Desse modo, podemos entender que os dados mostram que as regiões com características socioeconômicas e geográficas semelhantes foram afetadas de forma similar pelo coronavírus. Um caso típico é o de regiões com populações pequenas, que mostraram índices menores de incidência do vírus, possivelmente, por terem menor necessidade de se locomover e ter contato com outras pessoas. Já regiões populosas não conseguem se manter em isolamento e acabam aumentando os índices de casos e de óbitos.

REFERÊNCIAS

- ADMIN. Função para remover acentos no Excel. AMDS. Disponível em: <<https://amds.com.br/funcao-para-remover-acentos-no-excel/>>. Acesso em: 1 Jun. 2021.
- ALBUQUERQUE, C. et al. (2020) Apontamentos sobre a dimensão territorial da pandemia da COVID-19 e os fatores que contribuem para aumentar a vulnerabilidade socioespacial nas unidades e desenvolvimento humano de áreas metropolitanas brasileiras. A Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais [Dirur], nº 15, p. 7.
- ARGENTON, J. L. P. (2013). Árvore de Regressão para Dados Censurados e Correlacionados. Dissertação de Mestrado. UNICAMP – Instituto de Matemática, Estatística e Computação Científica.

BASTOS, S. B.; CAJUEIRO, D.O.; (2020) Modeling and forecasting the Covid-19 pandemic in Brazil. Departamento de Economia,FACE, Universidade de Brasília (UnB), Campus Universitário Darcy Ribeiro,National Institute of Science and Technology for Complex Systems (INCT-SC),Brasília, Brazil. Disponível em 01/04/2020 no endereço: <https://arxiv.org/abs/2003.14288>

BENARRÓS, G.; (2020). KCI's: Key Corona Indicators. Disponível em 29/03/2020 no endereço: <https://medium.com/@benarros/kcis-key-corona-indicators-a1f430a14aed>

CABALLERO-ANTHONY, M. (2005). SARS in Asia: crisis, vulnerabilities, and regional responses. Asian Survey, v. 45, n. 3, p. 475-495, 2005.

CASTRO, L. et al. (2016). Introdução a Mineração de Dados: Conceitos básicos, algoritmos e aplicações. Ed. Saraiva.

CHOLLET, F. (2018). Deep Learning with Python. Manning Publications. 386 p

CORDEIRO, G. M.; DEMÉTRIO., C. G. B. (2013). Modelos Lineares Generalizados e Extensões. Departamento de Estatística e Informática, UFRPE; Departamento de Ciências Exatas, ESALQ, USP. Disponível em: <https://docs.ufpr.br/~taconeli/CE22517/LivClarice.pdf>

DAVILA, V.H.L. Análise de Regressão e Correlação. Departamento de Estatística. Universidade Estadual de Campinas (Unicamp - Imecc). Disponível em: <https://www.ime.unicamp.br/~hlachos/RegresCorr.pdf>

EATON, J.L. (1983). Children's temperament: cluster analysis of parent questionnaires. PhD Thesis. University of Florida. Gainesville, FL.

FILHO, L. M A.L (2019). Correlação e Regressão. Departamento de Estatística. Universidade Federal da Paraíba. Disponível em: <http://www.de.ufpb.br/~luiz/AED/Aula9.pdf>

GUIMARÃES, P.R.B (2018). Análise de Correlação e medidas de associação. Disponível em: <https://docs.ufpr.br/~jomarc/correlacao.pdf>

HAYKIN, S. (2009). Neural Networks and Learning Machines. 3ª ed, Editora Prentice-Hall, 2008

IBM - Documentação: Regression tree. Cognos Analytics 11.1x. Disponível em: <https://www.ibm.com/docs/pt-br/cognos-analytics/11.1.0?topic=tests-regression-tree>

JOHNSON, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241–254 (3684 citations in Google Scholar as de 1/1/2016). Disponível no endereço: <https://doi.org/10.1007/BF02289588>

KHALID, W.; HERBERT-HANSEN, Z. N. L. (2018). Using k-means clustering in international location decision. Journal of Global Operations and Strategic Sourcing. Disponível em: <https://doi.org/10.1108/JGOSS-11-2017-0056>

KOHONEN, T. (1982). "Self-organized formation of topological correct feature maps". Biological Cybernetics, 43, p. 59-69.

KOHONEN, T. (2013). "Essentials of the self-organizing map". Neural Networks 37, 5265p.

LIU, G. et al. (2018). Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. Beijing, China. Disponível em 14/02/2018 no endereço: <https://doi.org/10.1016/j.jclepro.2018.02.129>

MATTEUCCI, M. (2006). Tutorial in Clustering. Politecnico di Milano. Department of Electronics and Information. Milano, Italy. Disponível em 15/05/2006 no endereço: http://www.elet.polimi.it/upload/matteucci/Clustering/tutorial_html/

MINISTÉRIO DA SAÚDE. (2020). Sobre a Doença: O que é o Coronavírus? Disponível em 07/06/2020 no endereço: <https://coronavirus.saude.gov.br/sobre-a-doenca#o-que-e-covid>

MIOT, H. A (2018). Análise de correlação em estudos clínicos e experimentais. Scielo Brasil. Disponível em: <https://doi.org/10.1590/1677-5449.174118>

OCHI, L. et al. (2005). Clusterização em Mineração de Dados. Documento Técnico do Programa de Pós-Graduação em Computação. Instituto de Computação da Universidade Federal Fluminense. Niterói, RJ.

OLIVEIRA, B. (2019). O que são Modelos Lineares Generalizados? Operdata. Disponível em: <https://operdata.com.br/blog/o-que-sao-modelos-lineares-generalizados/>

ORGANIZAÇÃO MUNDIAL DE SAÚDE – OMS. World Healthy Organization, <https://www.who.int/>

O ESTADO DE S. PAULO (2020). CORONAVÍRUS MATA 400 MIL NO MUNDO. A7, Buenos Aires, 08 jun. 2020. Disponível em: <https://digital.estadao.com.br/o-estado-de-s-paulo/20200608/textview>

OUR WORLD IN DATA. Coronavírus (COVID-19) Testing. Disponível em: <https://ourworldindata.org/coronavirus-testing>

PEDROSA, R.H.L.; (2020). The dynamics of Covid-19: weather, demographics and infection timeline. State University of Campinas (Unicamp), Campinas, Brazil. Disponível em 10/05/2020 no endereço: <https://doi.org/10.1101/2020.04.21.20074450>

R Documentation. <https://www.r-project.org/>. Acesso em: Jan 2020

RESTA, M.; SONNESSA, M.; TÀNGANI, E.; TESTI, A. (2017). Unsupervised neural networks for clustering emergent patient flows. Department of Economics and Business Studies, University of Genova, Via Vivaldi 5, 16126 Genova, Italy. Disponível em 26/08/2017 no endereço: <https://doi.org/10.1016/j.orhc.2017.08.002>

SILVA, L. et al. (2016). “Introdução à Mineração de Dados: Com Aplicações em R”. 1ª Ed. Elsevier Editora Ltda. R.J. 277 p.

SILVA, P. L. N.; (2020). Pandemia, pesquisa e estatísticas públicas: desafios e perspectivas. Escola Nacional de Ciências Estatísticas (Ence/IBGE) e pesquisador da Sociedade para o Desenvolvimento da Pesquisa Científica (SCIENCE). Disponível em 06/05/2020 no endereço: https://www.ipea.gov.br/coronavirus/assets/img/200506_seminario_ipea.pdf

ZHAO, Yanchang (2015). R and Data Mining: Examples and Case Studies. R and Data Mining. Disponível em: <http://www.rdatamining.com/docs/r-and-data-mining-examples-and-case-studies>. Acesso em maio de 2020.

Contatos: dani.deutner@hotmail.com e arnaldo.vallim@mackenzie.br