2. Coleta de Dados

Os dados dos municípios metropolitanos de SP, população, taxas de isolamento, quantidade de leitos UTI nos hospitais públicos e privados, geolocalização, número de casos confirmados e mortos pela COVID-19 são reais e foram obtidos através de sites disponibilizados pelo Governo do Estado de São Paulo.

2.1 Conjuntos de dados

Para analisar o problema utilizamos os seguintes conjuntos de dados (datasets) que estão no formato CSV (Comma-Separated Values):

* municipios\_rmsp.csv: Contém os 39 municípios da região metropolitana de São Paulo com a geolocalização (Fonte: EMPLASA - Empresa Paulista de Planejamento Metropolitano S/A)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| Município | String | Nome do município |
| Região | String | Região do município |
| Latitude | Number (double) | Latitude do município |
| longitude | Number (double) | Longitude do munícipio |

* covid19-cidades-sp.csv: Contém uma média de 2984 ocorrências diárias dos casos de COVID-19 dos municípios do estado de São Paulo (Fonte: Brasil.IO - O Brasil em dados libertos)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| Date | String | Data da ocorrência |
| State | String | Sigla do estado |
| City | String | Nome do município |
| Place\_type | String | Classificação da origem (cidade = city ou estado = state) |
| Confirmed | Number (integer) | Número de casos confirmados de COVID19 |
| Deaths | Number (integer) | Número de mortos de COVID19 |
| Is\_last | String | Indicador de atualização do dado |
| Estimated\_population\_2019 | Number(integer) | População estimada em 2019 |
| City\_ibge | Number (integer) | Código identificador da cidade no IBGE |
| Confirmed\_per\_100K\_inhabitants | Number (double) | Número de confirmados de COVID19 para cada 100.000 habitantes |
| Death\_rate | Number (double) | Taxa de mortalidade |

* dados\_populacao\_rmsp.csv: Contém alguns dados dos 38 municípios, tais como: área, população, índices de envelhecimento, grau de urbanização, densidade geográfica dos municípios do estado São Paulo (Fonte: Fundação SEADE - Fundação Sistema Estadual de Análise de Dados)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| Área | String | Área territorial em Km^2 |
| População | String | Número de habitantes do munícipio / Km^2 |
| Densidade Demográfica | String | Densidade demográfica do município |
| Taxa Geométrica de Crescimento Anual da População – 2010/2020 | String | Taxa geométrica de crescimento anual da população (% a.a.) entre os anos 2010 a 2020 |
| Grau de Urbanização | String | Grau de urbanização em % |
| Índice de Envelhecimento | String | Índice de envelhecimento em % |
| População com Menos de 15 Anos | String | Número de habitantes com menos de 15 anos de idade em % |
| População com 60 Anos e mais | String | Número de habitantes com 60 anos e mais de idade em % |
| Razão de Sexos | String | Razão de sexos (masculino / feminino) em 2020 |
| município | String | Nome do município |

* leitos\_uti\_adulto\_covid19.csv: Contém uma média de 102 hospitais públicos e privados com leitos de UTI adulto disponibilizados pelo município para o tratamento da COVID-19 (Fonte: Ministério da Saúde - Cadastro Nacional de Estabelecimentos de Saúde)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| CNES | Number (integer) | Cadastro Nacional de Estabelecimento de Saúde |
| Estabelecimento | String | Nome do estabelecimento de saúde |
| Município | String | Nome do município |
| Qtd\_Leitos\_UTI\_Adulto | Number (integer) | Quantidade de leitos UTI adulto |
| SUS | Number (integer) | Quantidade de leitos UTI do SUS (Sistema Único de Saúde) |

* leitos\_uti\_pediatrica\_covid19.csv: Contém uma média de 17 hospitais públicos e privados com leitos de UTI pediátrica disponibilizados pelos municípios para o tratamento da COVID-19 (Fonte: Ministério da Saúde - Cadastro Nacional de Estabelecimentos de Saúde)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| CNES | Number (integer) | Cadastro Nacional de Estabelecimento de Saúde |
| Estabelecimento | String | Nome do estabelecimento de saúde |
| Município | String | Nome do município |
| Qtd\_Leitos\_UTI\_Pediatrica | Number (integer) | Quantidade de leitos UTI pediátrica |
| SUS\_Pediatrica | Number (integer) | Quantidade de leitos UTI pediátrica SUS (Sistema Único de Saúde) |

* taxa\_isolamento\_por\_municipio\_20042020.csv: Contém uma média de 4056 taxas diárias de isolamento por município até 20/04/2020 (Fonte: Sistema de Monitoramento Inteligente do Governo de São Paulo)

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| Cdufum | Number (integer) | Identificador do município |
| Município | String | Nome do município |
| DATA | String | Data da ocorrência |
| Número de registros | Number (integer) | Número de registro |
| População | Number (integer) | Número de habilitantes do munícipio |
| UF | String | Unidade Federativa (sigla do Estado) |
| Índice de isolamento | String | Índice de adesão ao isolamento em % |

2.1 Ferramentas

Para a coleta e tratamento de dados utilizamos a ferramenta KNIME.

KNIME é uma plataforma de integração, análise interativa de dados e relatórios, disponível sob os princípios de software livre e aberto. Ele permite a integração de vários componentes no campo de aprendizagem de máquina (Machine Learning), mineração de dados (Data Mining), pré-processamento de dados (ETL – Extraction Transformation Loading), ligação de várias fontes de dados e visualização de dados. Seu desenvolvimento iniciou-se em janeiro de 2004 na universidade de Konstanz, onde foi criado por uma equipe de programadores como um produto proprietário. Desde o início, o objetivo principal era criar uma plataforma de processamento de dados modular, altamente escalável e aberta, e permitir uma fácil integração de vários módulos para carregamento, processamento, transformação, análise e mineração visual de dados sem focar em nenhuma área específica de aplicação.

Em meados de 2006, a primeira versão pública da plataforma é lançada e desde então, algumas empresas farmacêuticas começaram a utilizar o KNIME. Outras áreas, tais como de análise de dados clientes (CRM – Customer Relationship Management), Business Intelligence (BI) e análise de dados financeiros também se beneficiam utilizando o KNIME como ferramenta de exploratória de dados.

Desde junho de 2008, a empresa KNIME AG com sede em Zurique (Suiça) fornece suporte técnico profissional e serviços de consultoria para a sua plataforma.

A partir de 2012, o KNIME tem sido usado por mais de 15.000 usuários reais (usuários que baixam regularmente atualizações) e não apenas na área de ciência naturais, mas também nas áreas de bancos, produção de automóveis, telecomunicações, editoras, empresas de consultoria e muitas outras indústrias e em grande número de grupos de pesquisas em todo o mundo.

No quadrante do Gartner de 2018, o KNIME aparece entre os 4 líderes da categoria “Data Science and Machine Learning Platforms”.



Figura 1 – Quadrante Gartner 2018

Para o processamento e aprendizado de máquina utilizamos o Python. (Matheus documentar sobre o Python, bibliotecas e algoritmos utilizados)

Para a visualização dos dados utilizamos o PowerBi. (Cris documentar sobre o PowerBi e apresentar os gráficos)

3. Processamento/Tratamento de Dados

Os datasets que contém os dados dos munícipios metropolitanos de SP e população foram consolidados considerando a coluna MUNICÍPIO.

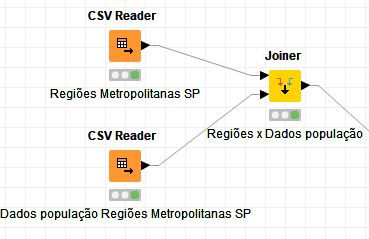


Figura 2 – KNIME Workflow | Agrupamento dados município com população

Para os datasets que contém dados dos leitos de UTI adulto e pediátrico foram necessários tratamentos antes de realizarmos o agrupamento desses dados.

O primeiro tratamento foi para os nomes dos municípios. Em ambos datasets, os nomes dos municípios estão em maiúsculo. Utilizamos um node (componente) chamado String Replace (Dictionary)) no qual para cada nome do município que estiver em maiúsculo, o mesmo será substituído pelo nome no padrão que constam nos demais datasets (primeira letra em maiúscula e os demais em minúsculo).

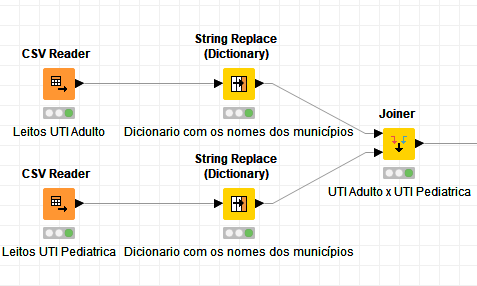


Figura 3 – KNIME Workflow | Agrupamento e tratamento dados leitos UTI

O segundo tratamento foi para os casos de municípios que possuem leitos UTI pediátrico e não possuem leitos UTI adultos, no qual foram encontrados no momento do agrupamento. Para não perder essa informação, recuperamos o nome do município e o estabelecimento que constam no dataset de Leitos UTI Pediatrica e os consideramos como dados principais.

Exemplo:

CNES: 2078325

Estabelecimento: HOSP NUM INFANTIL MENINO JESUS

Munícipio: São Paulo

O agrupamento dos dados de leitos da UTI adulto e pediátrico foram realizados considerando as colunas CNES e Município.

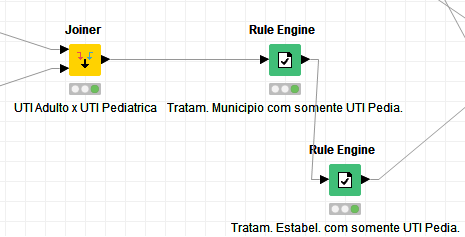


Figura 4 – KNIME Workflow | Agrupamento e tratamentos município e estabelecimento

Em seguida, realizamos os agrupamentos dos dados de Região e Leitos UTI Adulto Pediatrica (Região x UTIs) com os dados dos Casos de COVID-19 dos munícipios de SP. Para o agrupamento foi considerado a coluna MUNICIPIO.

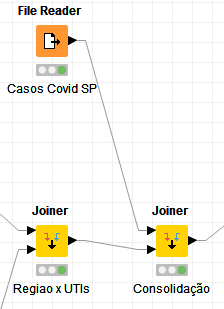


Figura 5 – KNIME Workflow | Agrupamentos Região x UTIs x Casos Covid SP

Após a consolidação, foram encontrados municípios de SP que possuem casos de COVID-19 tais como: Angatuba, Avaí, Bragança Paulista e Águas de São Pedro e não possuem dados de UTI Adulto e pediátrico. Como esses municípios não compreendem os munícipios metropolitanos de SP, realizamos o tratamento de eliminação desses dados.



Figura 6 – KNIME Workflow | Exclusões regiões que não possuem UTI

No dataset referente a Taxa de Isolamento consideramos as colunas MUNICIPIO, DATA e INDICE DE ISOLAMENTO, realizamos um tratamento no campo INDICE DE ISOLAMENTO para retirar o símbolo % e no campo DATA para o formato “yyyy-mm-dd” para equalização com o campo DATA do dataset de Casos COVID.

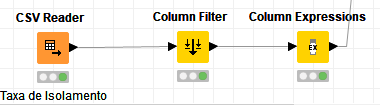


Figura 7 – KNIME Workflow | Taxa de Isolamento

Após os tratamentos realizados anteriormente, realizamos o agrupamento dos datasets de Casos COVID com Taxa de Isolamento utilizando as colunas DATA e MUNICIPIO.

Realizamos o tratamento para os valores ausentes ou nulos (missing values) de alguns campos tais como: CONFIRMADOS, MORTOS, INDICE DE ISOLAMENTO e RATE\_MORTOS.

As colunas QTD\_LEITOS\_UTI\_ADULTO, QTD\_LEITOS\_UTI\_PEDIATRICA\_SUS e SUS\_PEDIATRICA foram criadas para armazenar as somatórias das quantidades de leitos UTI adulto e pediátrica dos hospitais públicos e particulares.

Algumas colunas foram renomeadas para facilitar o entendimento tais como: CONFIRMED\_PER\_100K\_INHABITANTS para CONFIRMADOS\_POR\_100K\_HABIT, DEATH\_RATE para Rate\_Mortos e INDICE DE ISOLAMENTO para IND\_ISOLAMENTO(%).

Para as próximas etapas de análise e exploração dos dados, assim como na etapa de criação do modelo de Machine Learning, realizamos as conversões dos campos numéricos que utiliza a vírgula (,) como separador de casas decimais para ponto (.). Alguns campos que utilizam pontos tais como DENSIDADE\_DEMOGRAFICA e POPULACAO, realizamos a remoção dos pontos (.) para padronização.

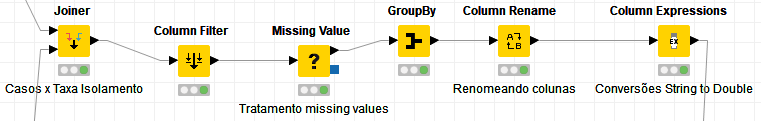


Figura 8 – KNIME Workflow | Agrupamento Casos x Taxas Isolamento e tratamentos

O arquivo final chamado dados\_consolidados\_regiao\_metropolitana\_sp.csv foi gerado contendo as seguintes colunas:



Figura 9 – KNIME Workflow | Arquivo dados\_consolidados\_regiao\_metropolitana\_sp

|  |  |  |
| --- | --- | --- |
| Coluna | Tipo | Descrição |
| Data | String | Data da ocorrência |
| Municipio | String | Nome do município |
| Confirmados | Number (integer) | Número de casos confirmados de COVID19 |
| Mortos | Number (integer) | Número de mortos de COVID19 |
| Cod\_IBGE | Number (integer) | Código identificador da cidade no IBGE |
| Confirmados\_Por\_100K\_Habit. | Number (double) | Número de confirmados de COVID19 para cada 100.000 habitantes |
| Rate\_Mortos | Number (double) | Taxa de mortalidade |
| Regiao | String | Região do município |
| Latitude | Number (double) | Latitude |
| Longitude | Number (double) | Longitude |
| Area | Number (double) | Área |
| Populacao | Number (double) | Número de habilitantes do munícipio |
| Densidade\_Geografica | Number (double) | Densidade demográfica do município |
| Tx\_Geomet\_Cresc\_Anual\_Popul\_2010\_2020 | Number (double) | Taxa geométrica de crescimento anual da população (% a.a.) entre os anos 2010 a 2020 |
| Grau\_Urbanizacao | Number (double) | Grau de urbanização em % |
| Ind\_Envelhecimento | Number (double) | Índice de envelhecimento em % |
| Popul\_Menos\_15Anos | Number (double) | Número de habitantes com menos de 15 anos de idade em % |
| Popul\_60Anos\_Mais | Number (double) | Número de habitantes com 60 anos e mais de idade em % |
| Razao\_Sexos | Number (double) | Razão de sexos (masculino / feminino) em 2020 |
| Ind\_Isolamento(%) | Number (double) | Índice de adesão ao isolamento em % |
| Total\_Leitos\_UTI\_Adulto | Number (integer) | Total de leitos UTI adulto do município |
| Total\_Leitos\_UTI\_Pediatrica | Number (integer) | Total de leitos UTI pediátrica do munícipio |
| Total\_SUS | Number (integer) | Total de leitos SUS |
| Total\_SUS\_Pediatrica | Number (integer) | Total de leitos SUS pediátrica |

A visão geral do Workflow no KNIME referente a coleta e tratamento dos dados:

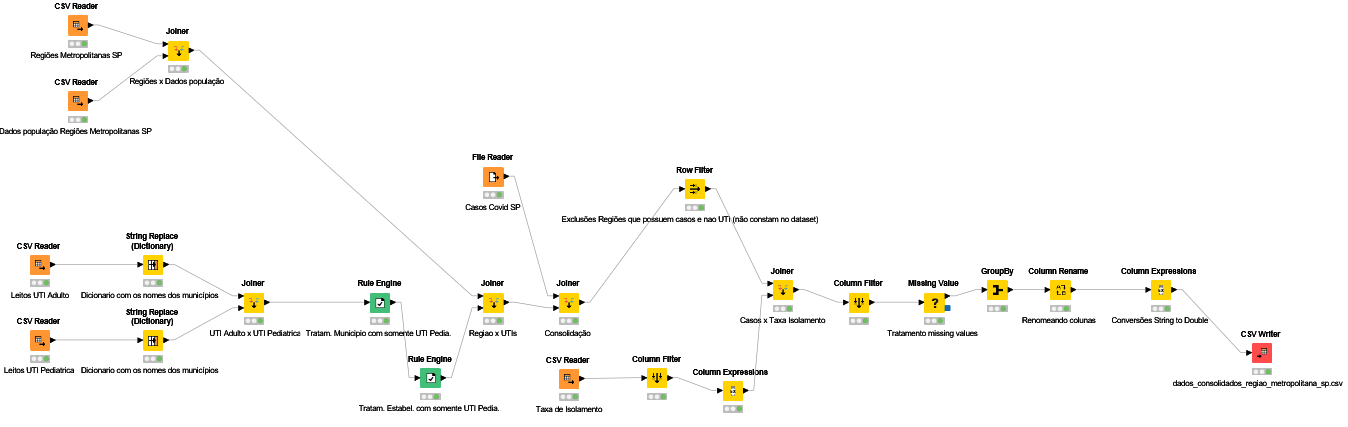


Figura 10 – KNIME Workflow | Visão Geral

X. Referências

EMPLASA, Empresa Paulista de Planejamento Metropolitano S/A. Arquivo CSV criado através dos dados disponíveis diretamente no site. Disponível em: <<https://emplasa.sp.gov.br/>>. Acesso em: 18 abr. 2020.

BRASIL IO, O Brasil em dados libertos. Disponibilização do arquivo no formato CSV dos casos de COVID-19 em todo o país. Disponível em: <<https://brasil.io/covid19/>>. Acesso em: 18 abr. 2020.

SEADE, Fundação Sistema Estatual de Análise de Dados. Disponível em: <<https://www.seade.gov.br/>>. Acesso em: 18 abr. 2020. Arquivo CSV gerado através da criação da API Set Dataset ([API\_Set\_Dataset.ipynb](file:///D:\Treinamentos\PUC-Minas-CienciaDados-BigData\13%20-%20TCC%20-%20Ciência%20de%20Dados%20e%20Big%20Data%20(2018)\TCC\API_Set_Dataset.ipynb) - notebook) disponível em: <<https://github.com/matheuxrp/TCC-PucMinas>>

MINISTERIO DA SAUDE, Cadastro Nacional de Estabelecimentos de Saúde, arquivo leitos\_UTI\_adulto\_covid19.csv criado através dos dados disponíveis diretamente no site. Disponível em: < http://cnes2.datasus.gov.br/Mod\_Ind\_Leitos\_Listar.asp?VCod\_Leito=51&VTipo\_Leito=3&VListar=1&VEstado=35&VMun=&VComp=>. Acesso em: 18 abr. 2020.

MINISTERIO DA SAUDE, Cadastro Nacional de Estabelecimentos de Saúde, arquivo leitos\_UTI\_pediatrica\_covid19.csv criado através dos dados disponíveis diretamente no site. Disponível em: < http://cnes2.datasus.gov.br/Mod\_Ind\_Leitos\_Listar.asp?VCod\_Leito=52&VTipo\_Leito=3&VListar=1&VEstado=35&VMun=&VComp=>. Acesso em: 18 abr. 2020.

SISTEMA MONITORAMENTO INTELIGENTE DO GOVERNO DE SÃO PAULO, adesão ao isolamento social em SP. Dados disponibilizados através da ferramenta de visualização TABLEAU. Disponível em: <<https://www.saopaulo.sp.gov.br/coronavirus/isolamento>>. Acesso em: 20 abr. 2020.

KNIME, artigo. Disponível em: <<https://pl.wikipedia.org/wiki/KNIME>>. Acesso em: 27 abr. 2020.

KNIME, Gartner Leader in Data Science Article. Disponível em: <<https://www.knime.com/about/news/knime-recognized-by-gartner-as-a-leader-in-data-science-and-machine-learning-platforms-2019>>. Acesso em: 27 abr. 2020.

==🡺 DICAS PARA REFERENCIAS 🡸

**Como escrever na lista de referências:**  
  
TYBEL, D. *Guia da Monografia*, 2016. Disponível em: <http://guiadamonografia.com.br/>. Acesso em: 25 dez. 2017.  
  
SIEMENS, G. *Connectivism*:    a    learning    theory    for    the    digital    age. 2004. Disponível em: <http://www.elearnspace.org/Articles/connectivism.htm>. Acesso em: 23 jan. 2016.  
  
(SOBRENOME, PRENOME abreviado do autor. *Título*: subtítulo (se houver) e data da publicação. "Disponível em:". "Acesso em:", seguido da data, exatamente como está nos exemplos.)  
  
LOPES, M. T. Inovação curricular no ensino superior. *Revista e-curriculum*, São Paulo, v. 7, n. 2, p. 45-70, 2011. Disponível em:  <http://revistas.pucsp.br/index.php/curriculum/article/ view/6852/4966>. Acesso em: 8 set. 2017.  
  
(Para textos impressos seguir as normas já estabelecidas e acrescentar os dados da internet, incluindo o endereço e data de acesso.)