Pontifícia Universidade Católica de Minas Gerais - PUC/MG

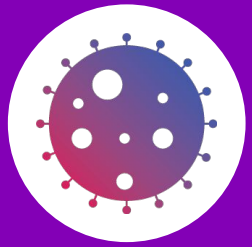
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

PUC Minas

Covid-19 no Estado de Minas Gerais: Uma análise do padrão de propagação e a influência de fatores no contexto da pandemia do novo Coronavírus.

Leonardo Alves Mateus

Contextualização



- Dezembro/2019: China revela surto de novo vírus identificado como **SARS-Cov-2**
- Março/2020: a **primeira morte** por COVID-19 No Brasil
- Novembro/2020: **160 mil mortes** e aproximadamente **5,4 milhões de infectados**.

- **Minas Gerais** é o maior estado do sudeste
- **586.528 km²** de extensão territorial
- **21 milhões** de habitantes aproximadamente
- Possui **853 municípios**

Problema Proposto

Compreender os diversos cenários de propagação do Coronavírus nos 853 municípios de Minas Gerais, através da **clusterização de séries temporais**, bem como verificar a **correlação entre variáveis demográficas** para com os diferentes padrões de disseminação da doença.

Técnica 5-Ws

Why?

Compreender os padrões de propagação da pandemia nos municípios mineiros e analisar os possíveis fatores correlatos.

Who?

Dados da Secretaria Estadual de Saúde de Minas Gerais (SES/MG), além de dados abertos do IBGE e SUS/Ministério da Saúde

What?

Agrupar os padrões de propagação, através de técnicas de Machine Learning, e analisar possíveis correlações.

When?

O período de análise compreende os estágios iniciais da pandemia no estado, a partir de **01/03/2020**, e se estende até a linha de corte para este trabalho, em **18/11/2020**.

Where?

Nos 853 municípios do estado de Minas Gerais

Coleta de Dados

- Dados de diferentes fontes governamentais, disponibilizados em sítios oficiais.
- Datasets foram obtidos em geral nos formatos CSV e XLSX
- A biblioteca utilizada para leitura foi o Pandas - ***pandas.read_csv()***
- Os dados foram armazenadas em dataframes - ***pandas.DataFrame()***

Datasets

- **CSV_Painel** (SES/MG-2020)
  - **CSV_Painel** – Confirmados
  - **CSV_Painel** – Óbitos
- **CSV_Sistemas** (SES/MG-2020)
- **IBGE Cidades e Estados** (IBGE)

Processamento/Tratamento de Dados

- Conversão de datasets - Excel (XLSX) para o formato CSV
- Tratamento específico para tags HTML existentes em alguns Datasets
- Valores ausentes preenchidos com zero ***Numpy.fillna(0)***, para não afetar as totalizações.
- Interligação dos Datasets utilizando o campo (chave) de referência **COD_IBGE**

Total de Registros

- **119.245** registros de casos confirmados
- **143.896** registros de óbitos
- **853** registros do IBGE

Análise e Exploração dos Dados

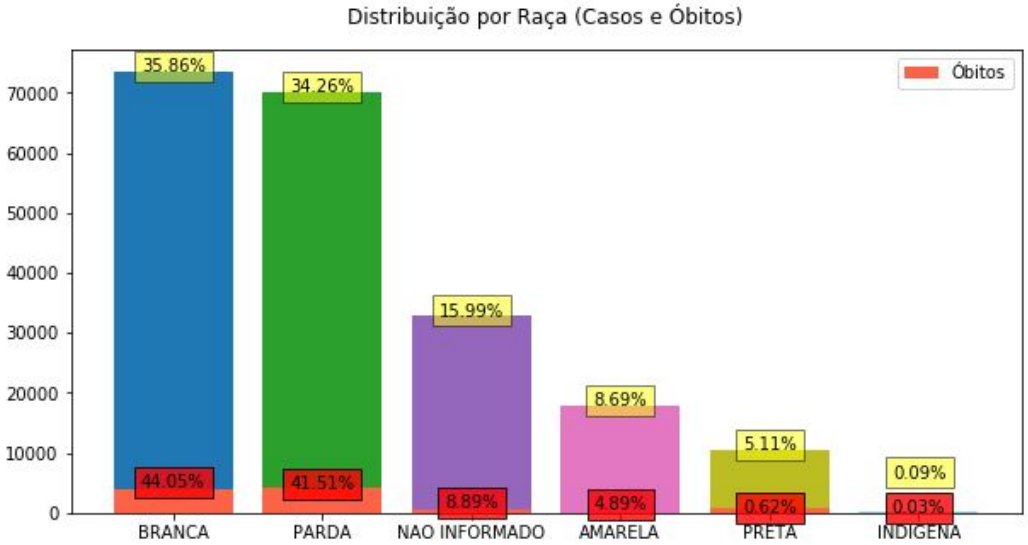
Análise Vertical

- Exploração de **dados qualitativos** da pandemia no Estado de Minas Gerais
- **Raça, Idade, Internação, Tipo de Internação, Comorbidade e Evolução do caso**
- Observação de como os dados estavam **distribuídos** entre as características
- **Obtenção de insights** sobre como a pandemia afetou a população do Estado de MG

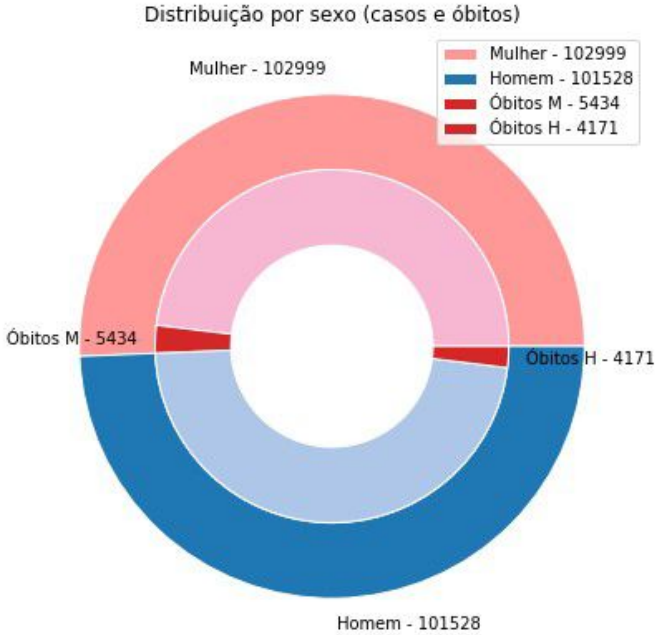
Análise Horizontal

- Foco em **séries temporais**
- Observação da **progressão do número de casos** e do acometimento de óbitos
- Análise da **velocidade** e o **alcance** da propagação da pandemia
- Exploração dos dados de **forma ampla** para **cada município** do estado

Análise Vertical

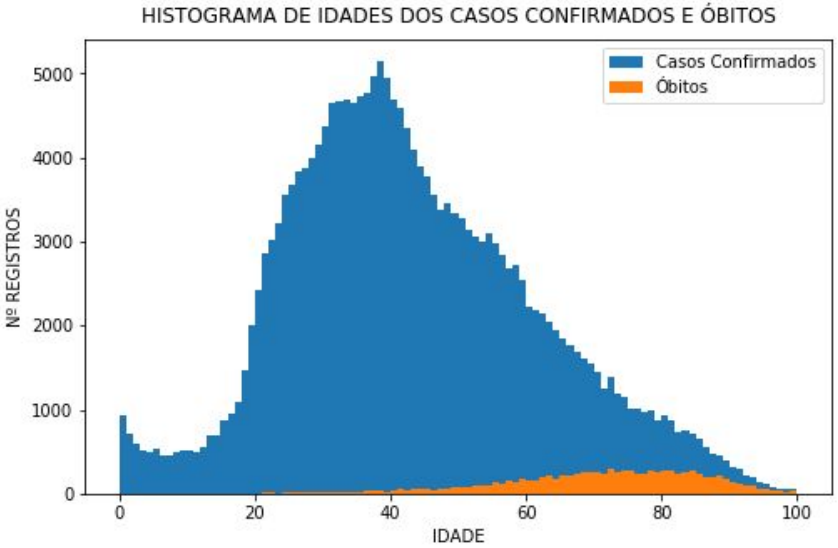


Predominância das Raças Branca e Parda

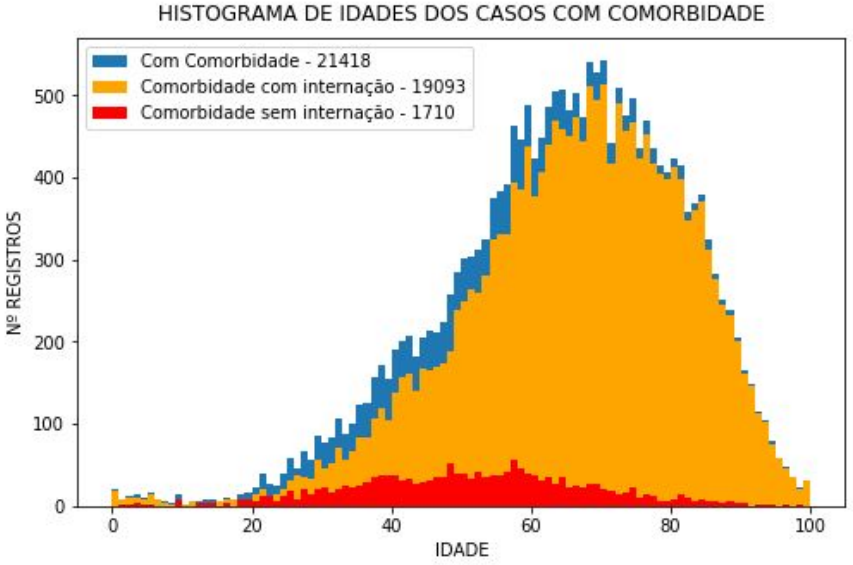


Disseminação Igualitária entre Sexos

Análise Vertical

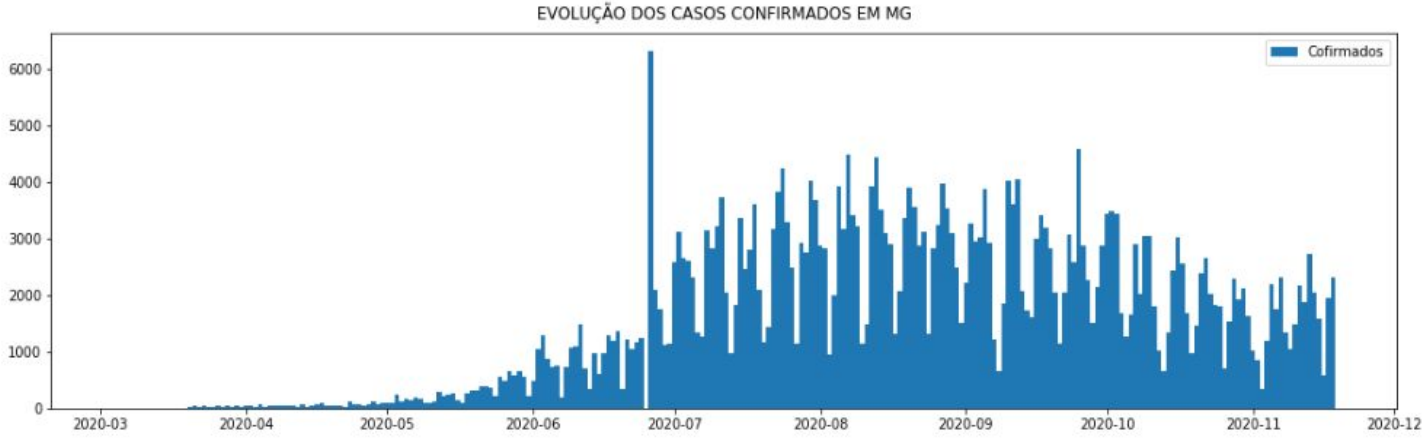


Concentração de casos em torno dos 40 anos, e de morte em torno dos 80 anos

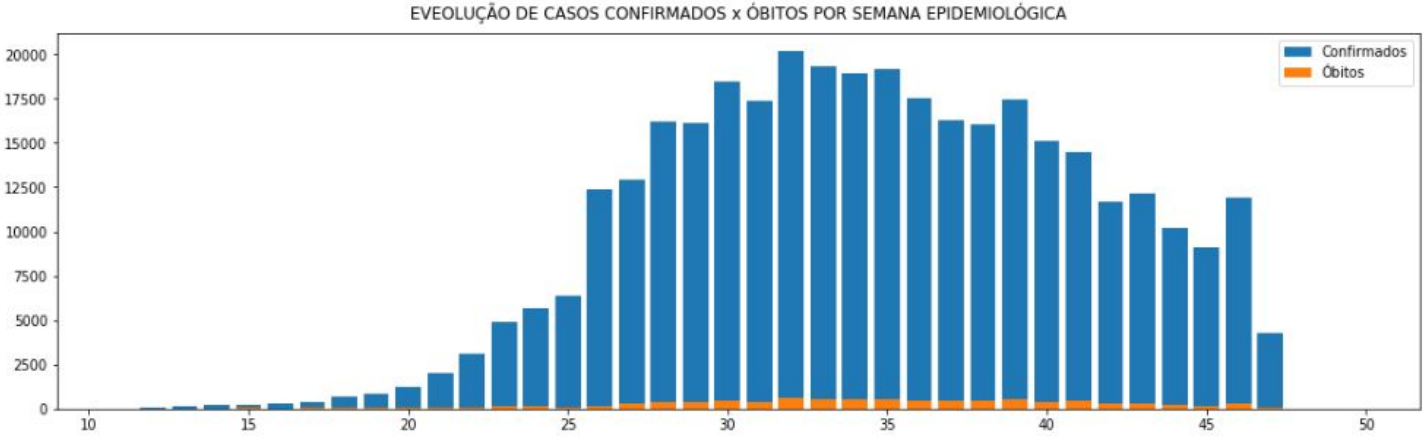


População com comorbidades entre 40 e 60 anos responderam melhor à doença, sem internações

Análise Horizontal



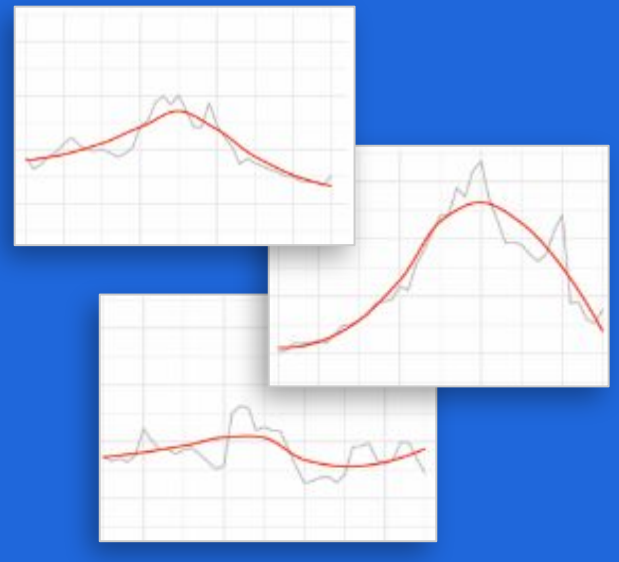
Análise Horizontal



Modelos de Machine Learning

Clusterização de Séries Temporais

- Algoritmo **K-Means**
- Aprendizado não supervisionado
- Agrupamento de séries temporais de propagação da pandemia em clusters representativos
- Representação de características como amplitude alcançada em números de casos
- Traçado da **linha de tendência** através da semanas e os **aspectos das curvas**



Modelos de Machine Learning

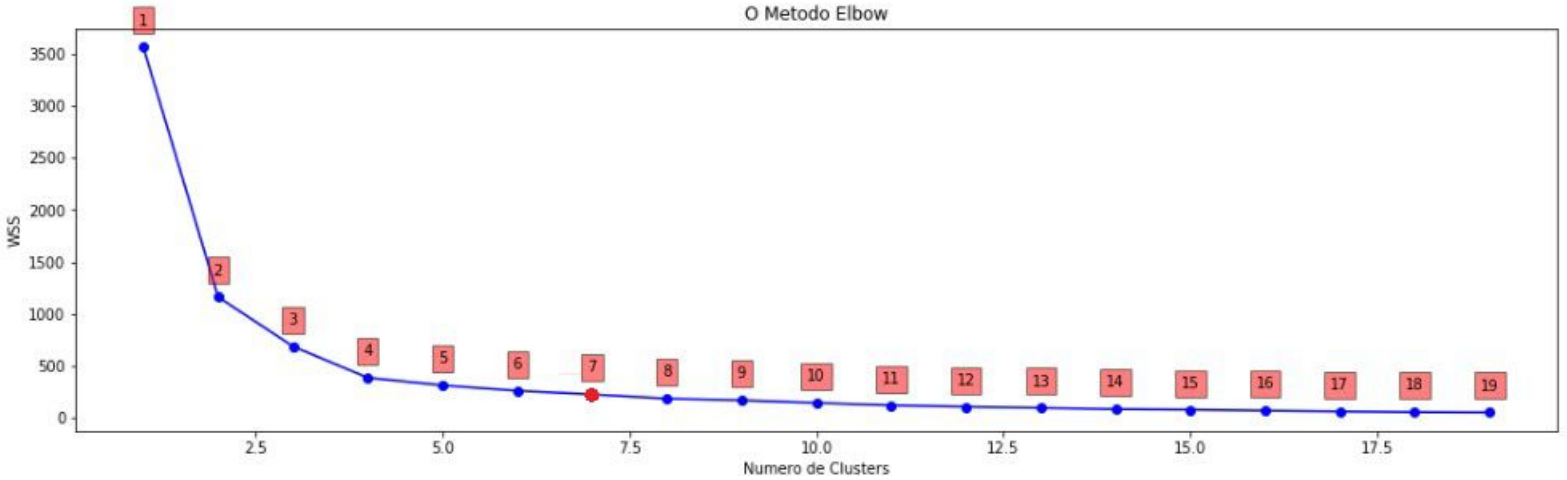
Formatação do Data Frame

	MUNICIPIO	S10	S11	S12	S13	S14	S15	S16	S17	S18	...	S38	S39	S40	S41	S42	S43	S44	S45	S46	S47
0	IPATINGA	0	0	0	0	0	2	2	5	2	...	250	223	176	178	235	307	221	234	391	87
1	DIVINOPOLIS	0	0	0	5	4	7	8	17	19	...	108	83	81	85	74	149	103	86	107	63
2	JUIZ DE FORA	0	0	5	5	21	17	20	21	41	...	371	308	226	190	229	268	155	259	575	177
3	PATROCINIO	0	0	0	0	2	1	1	0	0	...	148	125	120	79	110	69	29	15	66	3
4	BELO HORIZONTE	0	0	52	78	76	83	89	81	273	...	1734	1994	1567	1482	1323	1179	1202	990	976	907
5	CORONEL FABRICIANO	0	0	0	0	0	0	3	0	0	...	95	112	85	111	74	86	108	86	244	62
6	NOVA LIMA	0	0	6	11	13	6	0	2	6	...	82	135	167	216	460	17	132	93	145	158
7	SETE LAGOAS	0	0	0	0	0	0	2	0	0	...	213	229	109	131	84	106	51	82	105	79
8	UBERLANDIA	0	0	2	5	18	10	10	24	44	...	1327	1748	1706	1279	1296	985	795	461	547	151
9	MARIANA	0	0	0	0	0	0	3	5	1	...	43	75	45	58	41	56	47	3	59	0

Municípios na horizontal (linhas) e
Contagem de casos por semana na vertical (colunas)

Modelos de Machine Learning

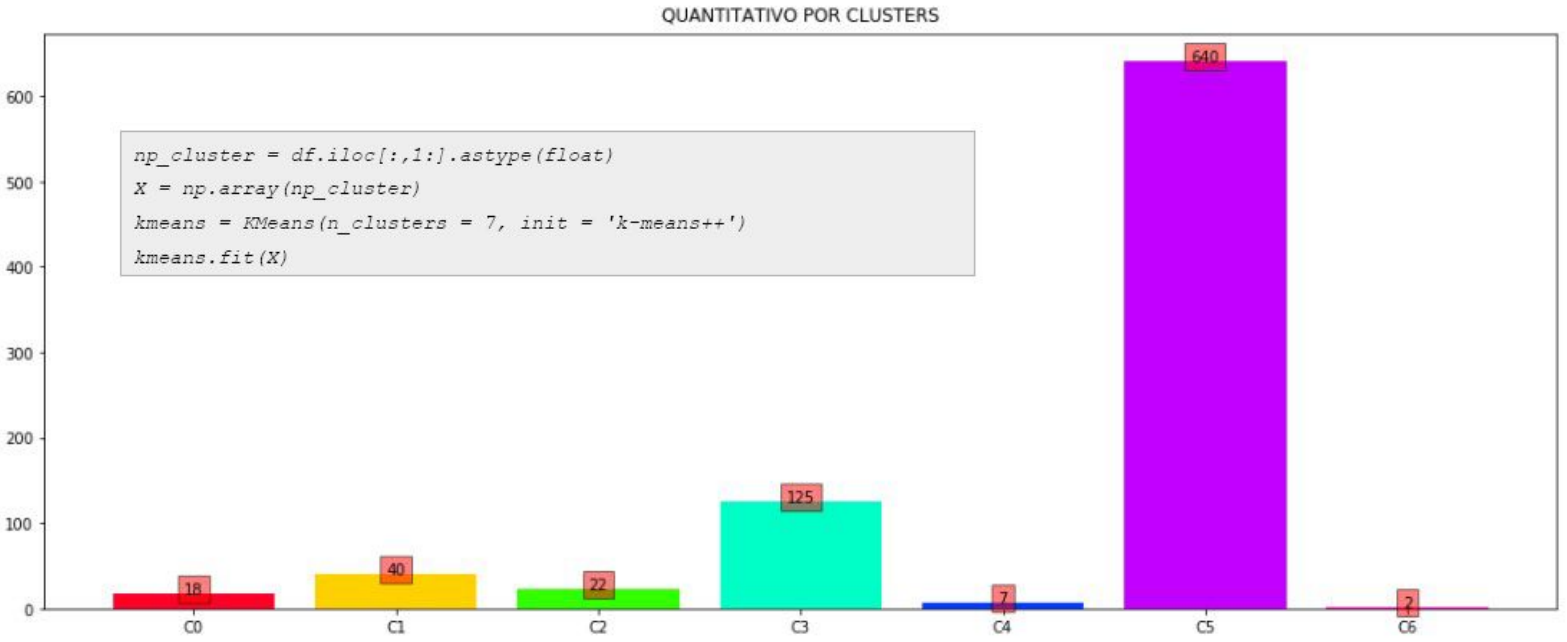
Número ideal de clusters: Método Elbow (Cotovelo)



Número escolhido: 7 Clusters

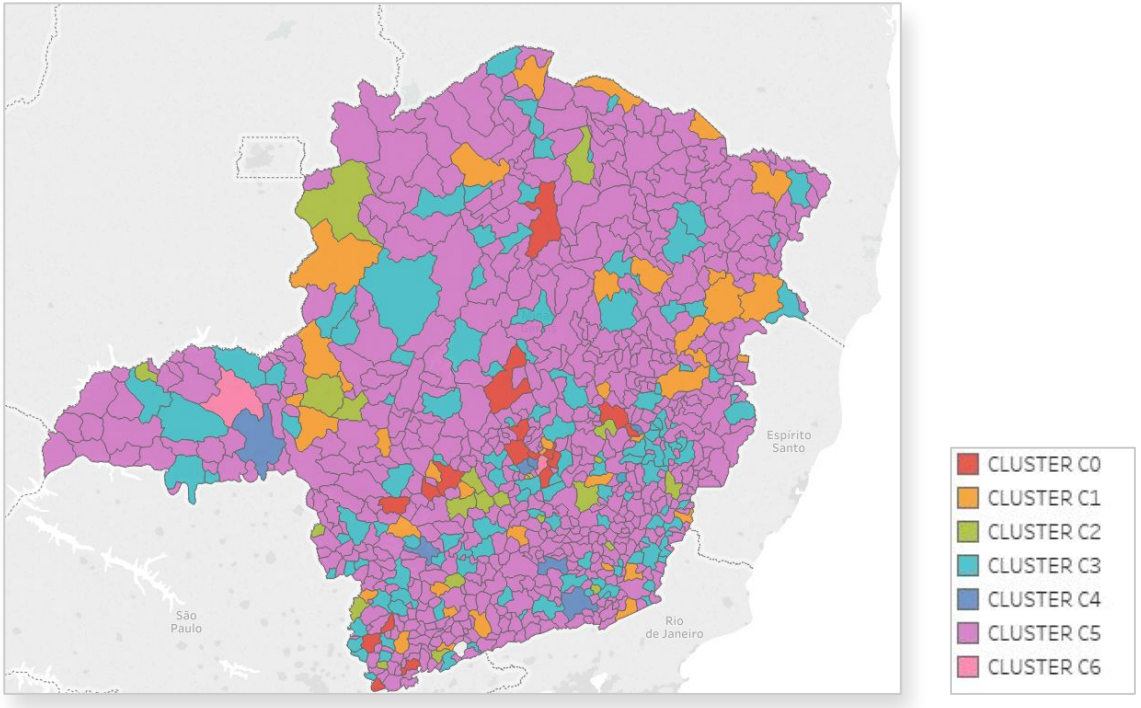
Modelos de Machine Learning

Resultado da execução: Quantitativo por Clusters



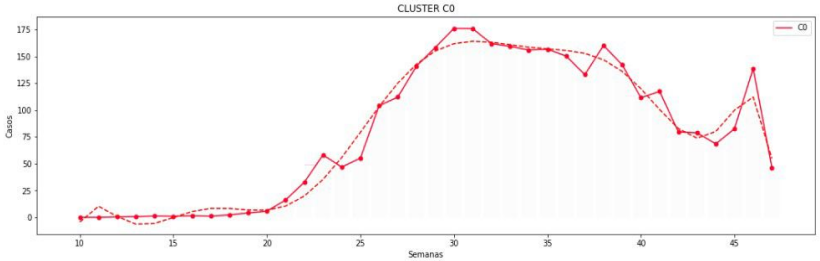
Modelos de Machine Learning

Resultado da execução: Mapa dos Clusters



Apresentação dos Resultados

Análise de Clusters

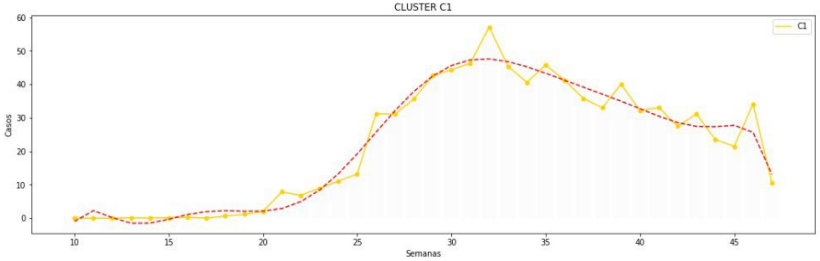


Cluster C0

Agrupamento: 18 municípios

Início da propagação: 20ª semana (10/maio)

Pico: 175 casos na 30ª semana (19/julho)

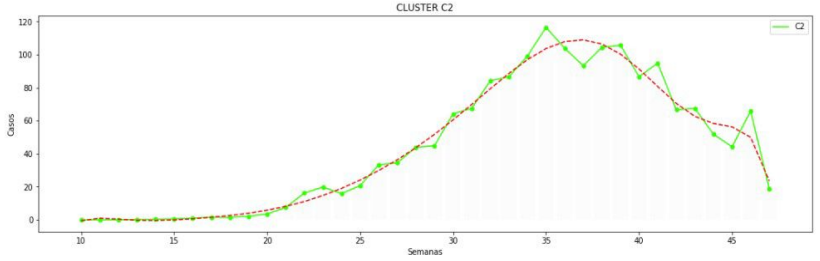


Cluster C1

Agrupamento: 40 municípios

Início da propagação: 22ª semana (24/maio)

Pico: 60 casos na 30ª semana (19/julho)



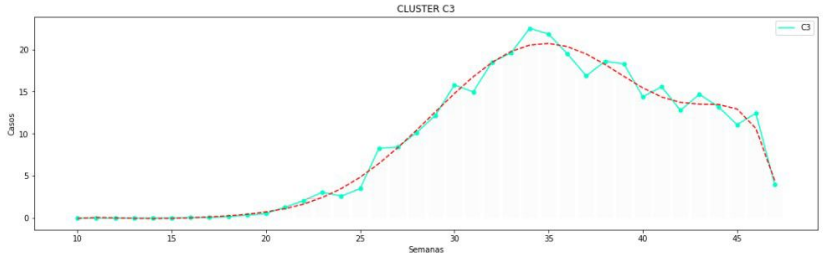
Cluster C2

Agrupamento: 22 municípios

Início da propagação: 25ª semana (14/junho)

Pico: 110 casos na 37ª semana (6/setembro)

Análise de Clusters

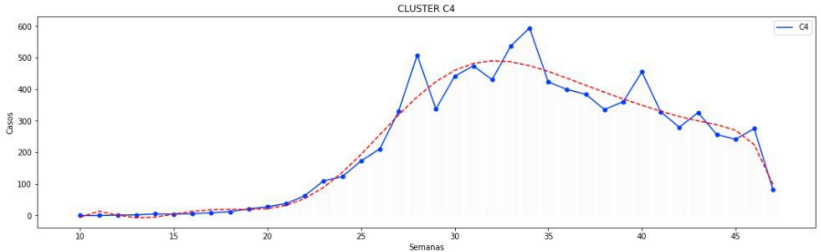


Cluster C3

Agrupamento: 125 municípios

Início da propagação: 25ª semana (14/junho)

Pico: 30 casos na semana 34ª (16/agosto)



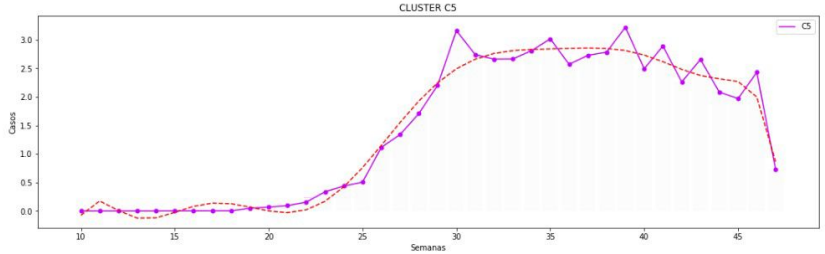
Cluster C4

Agrupamento: 7 municípios

Início da propagação: 21ª semana (17/maio)

Pico: 800 casos na 34 semana (16/agosto)

Análise de Clusters

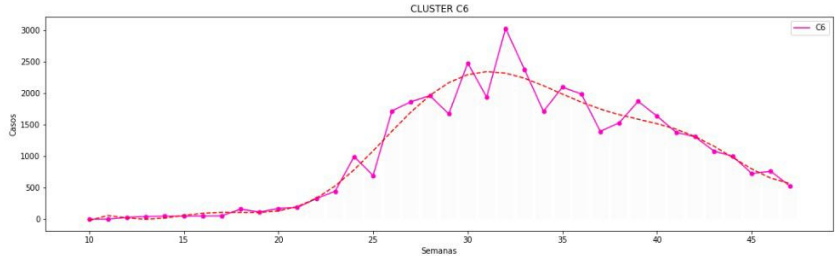


Cluster C5

Agrupamento: 640 municípios

Início da propagação: 25ª semana (14/junho)

Pico: 3 casos semana 34 (16/agosto)



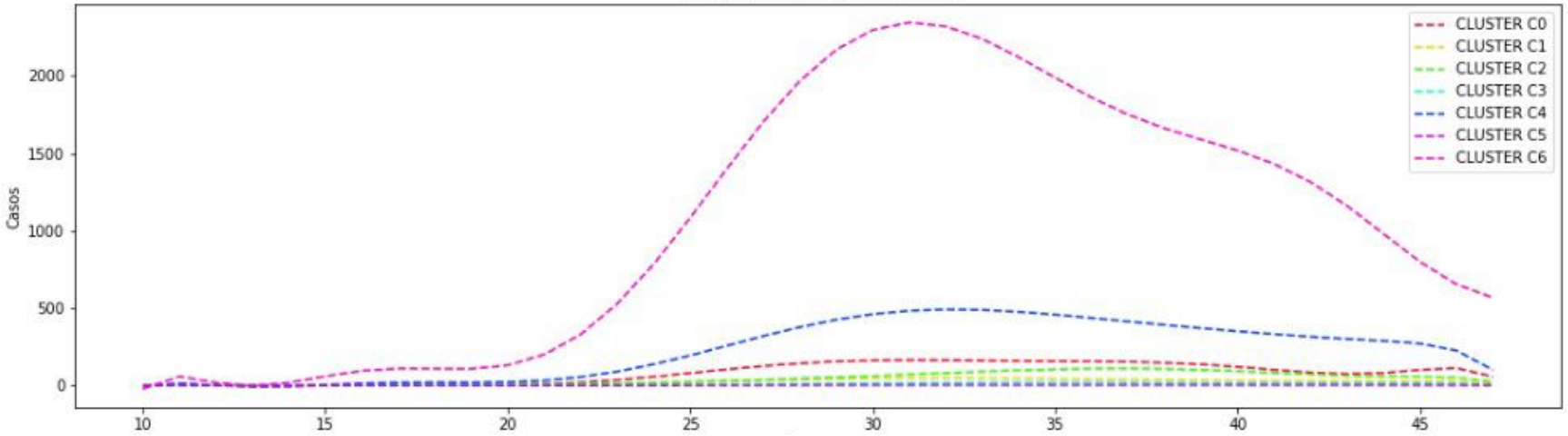
Cluster C6

Agrupamento: 2 municípios

Início da propagação: 18ª semana (16/abril)

Pico: 3000 casos na 32ª semana (02/agosto)

Resultado da Clusterização



Todos os clusters em uma mesma escala

Métrica dos Clusters

Coeficiente de Silhueta

Tabela referência

Coeficiente de Silhueta	Interpretação
0,71 a 1,00	Estrutura forte
0,51 a 0,70	**Estrutura razoável**
0,26 a 0,50	Estrutura fraca
Menor que 0,25	Nenhuma estrutura

Fonte: Kaufman e Rousseeuw (1990)

$$s = \frac{b - a}{\max(a, b)}$$

Valores encontrados

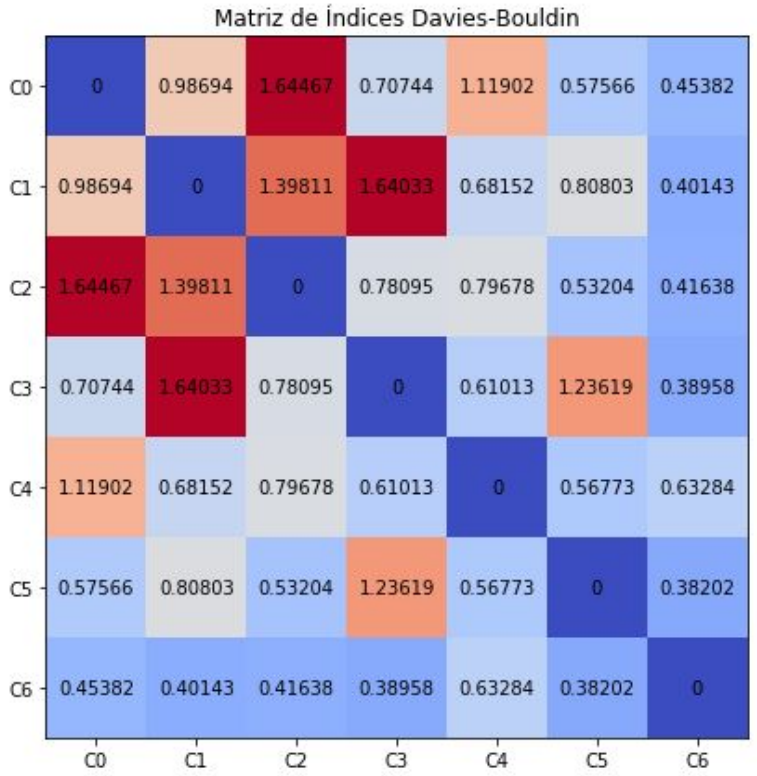
Cluster	Coeficiente de Silhueta
Cluster C0	-0.058691
Cluster C1	-0.035210
Cluster C2	0.099170
Cluster C3	-0.009606
Cluster C4	-0.003271
Cluster C5	0.687137
Cluster C6	0.065154
Valor Médio	**0.513339**

Métrica dos Clusters

Índice Davies-Bouldim

- Razão entre as distâncias "dentro do cluster" e "entre clusters".
- Valores mais próximos de zero indicam agrupamentos bem particionados.

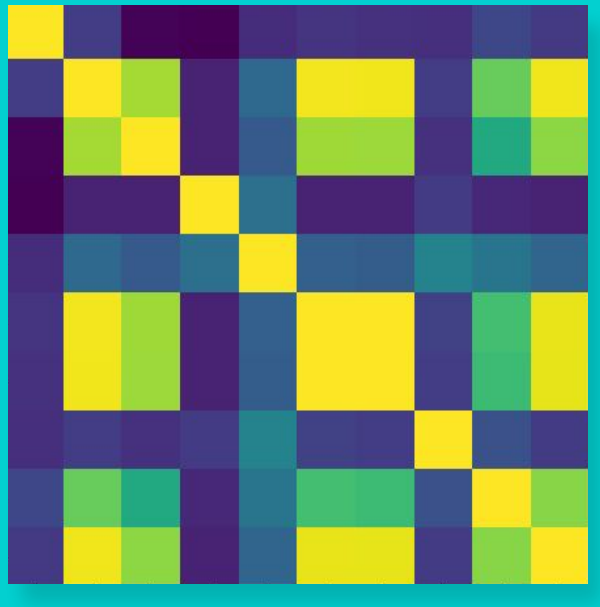
$$R_{i,j} = \frac{s_i + s_j}{d_{i,j}}$$



Análise de Correlação

Objetivos

- Confrontar cada grupo de municípios com seus fatores demográficos
- Identificar possíveis correlações entre estes fatores
- Avaliar a influência destes fatores na propagação das infecções e óbitos



Análise de Correlação

Fatores Demográficas *(IBGE)*

- Área Territorial
- População Estimada
- Densidade Demográfica
- Educação
- IDH Municipal
- Receitas realizadas no ano
- Despesas empenhadas no ano
- PIB per Capita

Variáveis Contextuais *(SES/MG)*

- Total de casos confirmados de Covid-19
- Total de óbitos por Covid-19

Análise de Correlação

Matriz de Correlação entre os Fatores Demográficas

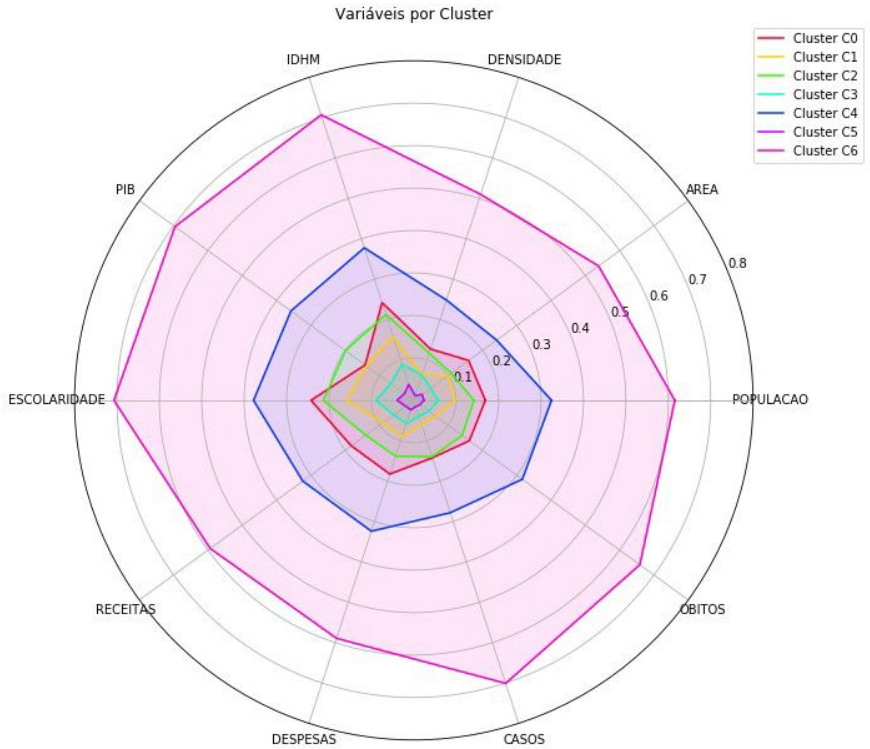
- Quanto mais populosos e densos os municípios, maiores as chances de propagação da doença e também maior o índice de mortalidade.
- Apesar de alguns locais possuírem grandes áreas, as cidades e locais de aglomeração de pessoas não são proporcionais às áreas territoriais.
- Os fatores IDHM, PIB per Capita e Índice de Escolaridade apresentaram graus de correlação baixos para influenciar o número de casos e óbitos

	AREA	POP	DENS	ESC	IDH	REC	DESP	PIB	CASOS	OBITOS
AREA_TERRITORIAL	1	0.114	-0.0732	-0.0854	0.057	0.078	0.0702	0.0614	0.149	0.0982
POPULACAO_ESTIMADA	0.114	1	0.854	0.021	0.285	0.981	0.977	0.11	0.749	0.976
DENSIDADE_DEMOGRAFICA	-0.0732	0.854	1	0.023	0.221	0.843	0.843	0.0713	0.575	0.813
ESCOLARIDADE	-0.0854	0.021	0.023	1	0.313	0.0233	0.0211	0.101	0.0383	0.0218
IDH	0.057	0.285	0.221	0.313	1	0.247	0.234	0.4	0.338	0.27
RECEITAS_REALIZADAS	0.078	0.981	0.843	0.0233	0.247	1	0.999	0.123	0.675	0.962
DESPESAS_EMPENHADAS	0.0702	0.977	0.843	0.0211	0.234	0.999	1	0.111	0.659	0.959
PIB_PERCAPITA	0.0614	0.11	0.0713	0.101	0.4	0.123	0.111	1	0.183	0.103
CASOS	0.149	0.749	0.575	0.0383	0.338	0.675	0.659	0.183	1	0.808
OBITOS	0.0982	0.976	0.813	0.0218	0.27	0.962	0.959	0.103	0.808	1

Alta correlação em **Vermelho**
Baixa correlação em **Azul**

Análise de Correlação

Matriz de Correlação entre os Clusters e os Fatores



- As Malhas Poligonais concêntricas evidenciam que a clusterização obedeceu também determinados padrões demográficos.
- O Cluster C6, de maior amplitude na curva de propagação, também detém maiores valores para as variáveis demográficas.
- O Cluster C5, que agrupou o maior número de municípios, apresenta as menores dimensões para as variáveis demográficas.

Conclusão

- Comparou os diferentes cenários de propagação da pandemia do Coronavírus para os 853 municípios de Minas Gerais
- Criou-se um modelo de aprendizado que dividiu as curvas de propagação da pandemia nos municípios em 7 cenários diferentes
- Associou-se os diferentes cenários a fatores como População, Área Territorial, Densidade Demográfica, IDHM, Receitas e Despesas
- Avaliou possíveis influências dos fatores na forma com que ocorreu a disseminação da doença.

PUC Minas

Obrigado!