PREVISÃO DA PROGRESSÃO DO COVID-19 E CLUSTERIZAÇÃO DE PAÍSES

Guilherme Lima Correa
FCA Unicamp
Limeira, Brasil
g173811@dac.unicamp.br

Resumo **- No ano de 2020 a sociedade global está enfrentando uma pandemia do novo corona vírus. O objetivo deste projeto é aplicar métodos de regressão afim de prever como o vírus pode se espalhar pelo mundo, ou seja, prever a progressão do vírus. Além disso, utilizar a clusterização para entendermos quais países estão no mesmo nível diante do vírus. Os dados foram retirados da plataforma que promove desafios de aprendizado de máquina chamada Kaggle.**

Palavras-chave: Corona Vírus 2019, Regressão, Clusterização.

1. Introdução

Atualmente a globalização repercutiu através de uma doença, isto é, a espécie humana como um todo enfrenta um desafio comum chamado corona vírus 2019. Este vírus compromete as vias respiratórias do indivíduo, podendo em casos extremos levar a morte. Neste cenário, há muitos estudos sendo feitos tanto por pesquisadores da iniciativa privada quanto pública, a fim de conhecer mais a doença. Um dos pontos analisados é a progressão do vírus, ou seja, a quantidade de pessoas que ele contamina diariamente.

Este estudo é importante para prevermos a devastação do vírus e assim tomarmos medidas preventivas a fim de evitar o cenário crítico. Neste trabalho foi utilizado dois métodos de regressão e uma de busca exaustiva para predição de casos contaminados. O primeiro, foi a regressão linear, em seguida, a regressão polinomial. E por último, utilizou-se um método denominado busca aleatória (*Random Search).* Ao final, foram comparadas algumas métricas para definir o melhor.

Além deste estudo sobre a progressão, nota-se no cotidiano comparações discrepantes entre países que possivelmente não poderiam ser relacionados. Por esse motivo, foi feito uma clusterização para avaliar quais países podem ser agrupados.

2. Banco de Dados

O conjunto de dados contém informações diárias a partir de a 22/01/2020 até 20/07/2020 sobre o número de casos confirmados, óbitos e recuperação do novo Corona vírus - 19. Observe que esses são dados de séries temporais e, portanto, o número de casos em um determinado dia é o número acumulado.

O arquivo principal deste conjunto de dados é covid_19_data.csv e as descrições detalhadas dos dados utilizados estão abaixo.

Dados	Descrição
Província / Estado	Província ou estado da observação (pode estar vazio quando estiver ausente)
País / região	País de observação
Confirmado	número acumulado de casos confirmados até essa data

Óbitos	Número acumulado de óbitos até essa data
Recuperado	Número acumulado de casos recuperados até essa data

Tabela 1: Dados obtidos da plataforma Kaggle

3. Regressão

3.1 Regressão Linear

Na regressão analisou-se o número de casos confirmados. O primeiro método testado foi a regressão linear. Dividiu-se o conjunto de dados entre teste e treino. Para teste foram utilizados 85% dos dados. Vale ressaltar, que por ser uma série temporal os dados não foram embaralhados.

A progressão do vírus não é linear, portanto, este método não foi capaz de prever progressão de maneira eficiente como podemos ver nas métricas.

A equação obtida através deste método é:

$$y = 58443.42 * x - 1918022.83$$

O erro médio absoluto (MAE) a métrica de qualidade mais básica que há para se analisar uma regressão, está é a soma de todos esses erros divido pelo número de pontos.

$$MAE = 40*10^5$$

O erro médio quadrático (MSE) tem como base o erro médio absoluto, contudo, o erro (distância entre os pontos e a reta) é elevado ao quadrado.

$$MSE= 1757*10^{10}$$

Ambos assumiram valores elevados, então optou-se por analisar também o Coeficiente de determinação R^2. Este assumiu um valor negativo: -5.49 que significa que o modelo escolhido foi pior do que uma linha horizontal. Portanto, o modelo escolhido não segue a tendência dos dados.

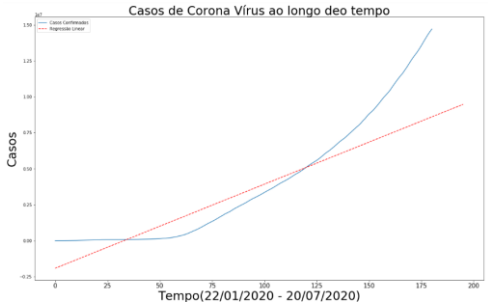

Figura 1: Regressão Linear para todo conjunto de dados.

3.2 Regressão Polinomial

O segundo método de regressão testado foi a polinomial, este se saiu melhor que o primeiro. Foram testados polinômios de diversas graus e o que obteve os melhores resultados foram o de terceiro.

$$MAE: 8*10^5$$

$$MSE: 89*10^{10}$$

Apesar de suas métrica mostrarem resultados melhores do que a regressão linear, podemos considerar que o resultado ainda não é satisfatório quando observamos apenas estes valores.

Porém, ao olhar o gráfico deste polinômio e o gráfico de casos confirmados no mundo, notamos uma semelhança:

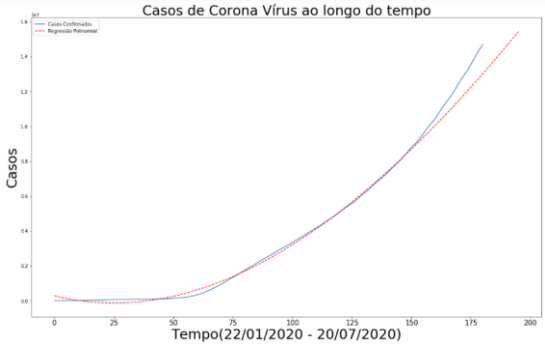

Figura 2: Regressão polinomial para todo o conjunto de dados

Nota-se que a curva obtida consegue explicar muito bem os dados de treino, porém para os dados de teste os erros passam a ser elevados.

Quando isso acontece, enfrentamos um problema de Overfitting — quando o modelo "adivinha" muito bem os dados que foram usados para treiná-lo, mas ele não consegue se sair muito bem com dados que nunca viu (teste).

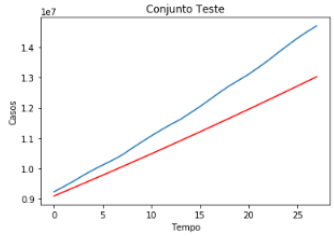

Figura 3: Regressão Polinomial para o conjunto de dados referente ao Teste

4. Busca Aleatória

É um método de busca exaustiva. Neste configuramos uma grade de valores de hiperparâmetros e seleciona-se combinações aleatórias para treinar o modelo. Isso permite controlar explicitamente o número de combinações de hiperparâmetros que são testadas, o que melhora o desempenho do algoritmo. O número de iterações a ser pesquisada é baseado no tempo ou recursos configurados.

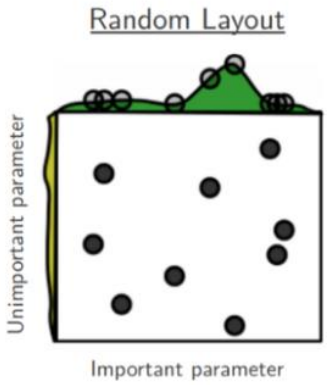

Figura 4: Funcionamento da Busca Aleatória

Para este método algumas métricas foram melhores do que as regressões testadas anteriormente.

MAE: $9*10^5$

MSE: $86*10^{10}$

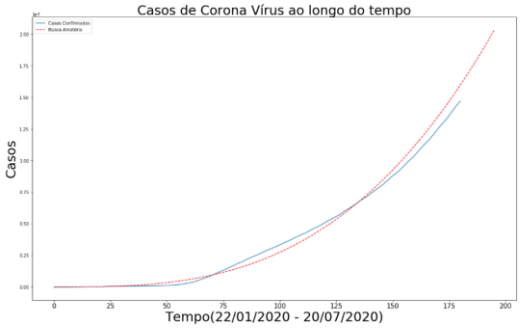

Figura 5: Busca Aleatória para todo conjunto de dados

Ao observar o gráfico para o conjunto de teste notamos que o erro se propagou de maneira menos evidente do que na regressão polinomial.

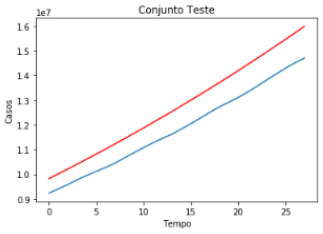

Figura 6: Busca Aleatória para o conjunto de dados referente ao Teste

5.Resultados da Regressão.

Devido à proximidade do modelo com os dados reais (principalmente do conjunto teste), foi realizada a previsão de casos confirmados para os próximos 15 dias tanto para a regressão polinomial quanto para a busca aleatória. Ou seja, do dia 21/07/2020 até 04/08/2020. Os dados que foram previstos foram comprados com os dados reais até o dia 29/07/2020.

Data	Casos Confirmados	Casos Confirmados (polinômio)	Casos Confirmados (BA)	Erro (polinômio)	Erro (BA)
07/21/2020	14.562.547	13.183.548	16.253.267	9,47%	11,61%
07/22/2020	14.765.253	13.340.443	16.523.814	9,65%	11,91%
07/23/2020	15.012.728	13.498.066	16.797.350	10,09%	11,89%
07/24/2020	15.296.919	13.656.413	17.073.892	10,72%	11,62%
07/25/2020	15.581.002	13.815.480	17.353.457	11,33%	11,38%
07/26/2020	16.055.909	13.975.264	17.636.060	12,96%	9,84%
07/27/2020	16.296.635	14.135.761	17.921.718	13,26%	9,97%
07/28/2020	16.737.842	14.296.969	18.210.448	14,58%	8,80%
07/29/2020	17.039.160	14.458.882	18.502.266	15,14%	8,59%

Tabela 2: Comparativo da Regressão Polinomial e B.A. com os dados reais do covid-19.

Podemos ver que o método de busca aleatória se saiu melhor diante dos métodos de regressão tradicionais. A média do percentual de erros de previsão foi de

6. Clusterização

Para isso foi necessário definir quais variáveis analisaríamos. Definiu-se duas variáveis implícitas no banco de dados que são a taxa de mortalidade e a taxa de recuperação.

$$tx\ de\ mortalidade = \frac{óbitos}{confirmados} \quad (1)$$

$$tx\ de\ recuperação = \frac{recuperação}{confirmados} \quad (2)$$

A partir disso, definiu-se que o conjunto de dados seria dividido em 6 cluster, isto através do *Método do Cotovelo*.

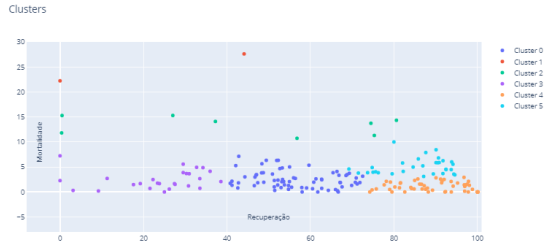

Figura 7: Clusters

Com a clusterização feita, analisou em qual cluster o Brasil se enquadrava.

Brasil	Irã	Alemanha	Canada
China	Polônia	Suíça	Irlanda
Japão	Áustria	Dinamarca	Finlândia
Cuba	Estônia	Uruguai	Cuba

Tabela 3: Países do Cluster 5

A partir disso analisou-se os dados de 5 países deste cluster incluindo o Brasil.

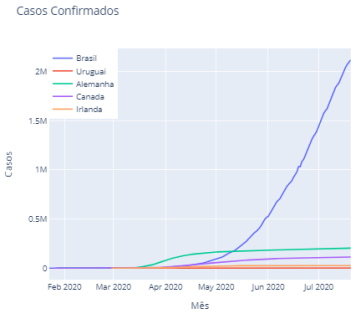

Figura 8: Análise de casos confirmados de países do Cluster 5

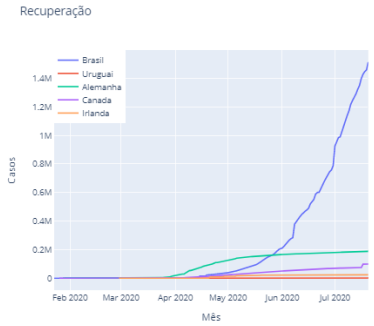

Figura 9: Análise de casos recuperados de países do Cluster 5

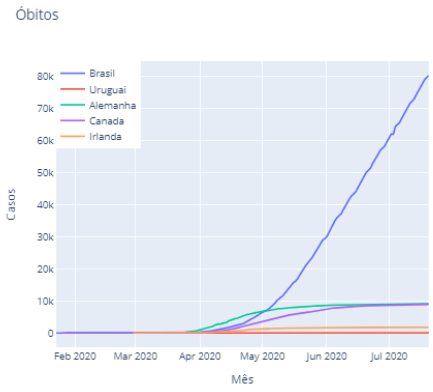

Figura 10: Análise de casos de óbitos de países do Cluster 5

Com estes gráficos é possível notar que a clusterização foi eficiente para alcançar o objetivo deste trabalho. O Brasil, por exemplo, se enquadra no Cluster ao qual o algoritmo o agrupou. Visto que, as três curvas traçadas não definem o cluster, já que este é definido através

das taxas de mortalidade e taxa de recuperação. Ou seja, independe do número de casos confirmados, pois quanto maior este número mais será o número de óbitos e recuperados. E como a taxa é uma razão entre essas grandezas, a quantidade de casos não foi um fator relevante para a clusterização.

PAÍSES	Tx de Mortalidade	Tx de Recuperação
'Brazil'	3.78	71.5
'Uruguay'	3.10	87.12
'Germany'	4.47	92.49
'Canada'	7.88	87.63
'Ireland'	6.80	90.677
Cluster 5	5.11	85.57

Tabela 4:Taxa de mortalidade e recuperação dos países do cluster 5

7.Conclusão

Como o trabalho está sendo realizado com dados recolhidos simultaneamente conforme a vírus avança, pode existir um erro de contagem, visto que, que um caso é confirmado apenas após o teste. Sabe-se que muitas pessoas não fazem os testes devido à falta de estrutura do próprio país.

O método de regressão não enquadrou a curva de progressão do vírus como uma exponencial, apensar de graficamente ela se assemelhar. Este método poderia ser implementado e, possivelmente, apresentaria resultados melhores do que a Busca Aleatória.

Por fim, na clusterização, pode-se fornecer mais dados para o modelo como PIB, renda per capita, número de habitante, entre outros fatores que divergem entre os países. A fim de captar mais a realidade com os dados para que o cluster forneça os melhores agrupamentos.

8.Referências Bibliográficas

TOMAZELI, Leonardo. Machine learning: Regressão e Clusterização . 01 mar. 2020, 30 jul. 2020. Notas de Aula.

Dados do corona vírus 2019. Disponível em: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv. Acesso em 26/07/2020

Dados do corona vírus 2019. Disponível em: https://news.google.com/covid19/map?hl=pt-BR&gl=BR&ceid=BR:pt-419. Acesso em 26/07/2020