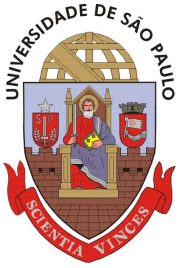

Universidade de São Paulo

SCC275 – Introdução de Ciências dos Dados

Análise dos novos casos semanais de COVID-19 em relação à Taxa de Isolamento Social no Estado de São Paulo

Alunos:

Eduardo Amaral (11735021)

Vivian Kiyomi Amano (11300502)

São Carlos, 14 de Dezembro de 2020

I - INTRODUÇÃO

A pandemia causada pelo patogênico SARS-CoV-2 impactou o mundo tanto na área da saúde quanto na econômica. A doença causada, COVID-19, tem como uma das principais características a alta taxa de contaminação, podendo causar uma série de síndromes respiratórias e em muitos casos ser letal (Briz-Redón & Serrano-Aroca, 2020). No mundo já foram contabilizados cerca de 67.027.780 casos, sendo o número de óbitos próximo a 1.535.492, registrados até o momento da publicação deste trabalho. No Brasil os números são também alarmantes. Cerca de 6.603.540 casos foram registrados, com 176.962 óbitos (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde, 2020).

Uma grande mobilização científica em diversas áreas vêm buscando analisar e compreender as principais características do vírus na tentativa de auxiliar os poderes públicos a tomarem as decisões cabíveis para conter a progressão da COVID-19 (Xu et al., 2020). Dessa forma, o presente trabalho busca construir duas bases de dados, uma da taxa de transmissão e progressão do vírus no Brasil e outra da taxa de adesão ao isolamento social, das principais cidades brasileiras. Podendo dessa forma, através de ferramentas estatísticas e modelagem de dados, analisar e entender as dependências entre a taxa de transmissão e adesão ao distanciamento social.

Os dados utilizados neste trabalho foram coletados por meio da API ElasticSearch do Open DataSUS, que fornece informações referentes aos casos de COVID-19 em todo Brasil. Outra fonte de dados utilizada foi o portal do estado de São Paulo que fornece dados referentes à taxa de isolamento social das cidades paulistas. Os dados foram pré-processados por meio de um script feito pelos alunos em Node.js. Durante o estudo foram utilizadas técnicas como a matriz de correlação de Pearson, para seleção de variáveis, o método PCA, para redução de dimensionalidade, análise de diversos classificadores em relação à uma métrica específica e curva ROC, para seleção do melhor classificador. Por fim observou-se que o melhor classificador em relação à maior parte das métricas foi o Multi-Layer Perceptron.

Contudo, além dessa seção que se refere a introdução e exposição da problemática e modo de abordagem, o presente trabalho é organizado em: Trabalhos relacionados, onde é descrito outros trabalhos correlacionados e como eles abordaram a mesma problemática; Materiais e métodos, que compreende a forma da construção e manipulação da dataset, assim apresenta a ferramenta e modelo utilizado para as análises; Experimentos e discussões, que expõem, através de tabelas e gráficos os resultados obtidos pela, discutindo sobre e fazendo comparações quando necessário; Conclusões, que sintetiza as constatações observadas sobre os dados, ferramenta e modelo empregado, levantado os seus pontos positivos e negativos.

II - TRABALHOS RELACIONADOS

1. The dynamics of Covid-19: weather, demographics and infection timeline

Neste artigo são estudado os efeitos dos fatores climáticos, densidade demográfica e locais na transmissão do vírus Covid-19 nos 50 estados do Estados Unidos e 110 países. O artigo em questão é referência do próximo artigo, pois as ferramentas utilizadas nele também foram usadas no outro.

Foram utilizado modelos lineares, através dos quais observou-se que apenas as variáveis de densidade populacional e da linha do tempo eram estatisticamente relevantes. Também concluiu-se que quanto maior a densidade demográfica, maior era a taxa diária inicial de crescimento do número de casos.

2. A dinâmica da Covid-19 no Brasil e em São Paulo: demografia, número básico de reprodução e o efeito do distanciamento social

Este artigo apresenta a relação do crescimento da Covid-19 nas capitais brasileiras e nas principais cidades do estado de São Paulo. Estimando o quanto as medidas de distanciamento social adotadas até o final de abril influenciaram na transmissão e no aumento de casos. Seus dados incluem fatores climáticos, linha do tempo da instalação da doença e densidade populacional. O estudo utilizou modelos lineares multivariáveis. Com base nesses dados, utilizou-se os moldes dos modelos do artigo anterior, e buscou-se resultados parecidos ao artigo base, pois, por mais que os modelos urbanos utilizados sejam distintos, os valores relevantes (como, taxas iniciais de crescimento da Covid-19) são próximos.

Por fim, são feitas estimativas para o Brasil e São Paulo com curvas de crescimento, e calcula-se números de atenuação que indicam os resultados do isolamento social em conjunto com outras medidas tomadas como uso de máscaras obrigatório em lugares públicos. Observa-se que tais números são alcançados com uma taxa de isolamento acima de 65%.

3. Medidas de distanciamento social e mobilidade na América do Sul durante a pandemia por COVID-19: Condições necessárias e suficientes?

Este artigo estuda o impacto das medidas de distanciamento social na América do Sul e o nível de transmissão da doença na comunidade. Inicia-se com duas hipóteses: a primeira é que em países que fizeram o lockdown há uma redução na mobilidade da população, e a segunda é que em alguns países selecionados (Argentina, Brasil e Colômbia) forma-se padrões de mobilidade. Por meio dos dados de geolocalização

disponibilizados pela Google, identifica-se as variações na circulação das pessoas na América do Sul. Para isso foi utilizado as tendências de mobilidade da população para um conjunto de países entre 15 de fevereiro e 16 de maio de 2020. Obtendo-se um indicador geral de circulação para identificar os padrões regionais que usa uma análise descritiva para a autocorrelação espacial .

Na análise da base de dados foi feito um pré-processamento para valores ausentes (substituindo por médias). A tendência de mobilidade foi uniformizada com uma única métrica e aplicou-se técnicas estatísticas (Season-Trend decomposition using LOESS). Visando comprovar a hipótese que a adesão ao distanciamento social não se atribui de maneira aleatória no espaço, e sim com formações de padrões regionais de mobilidade, utilizou-se uma análise de autocorrelação espacial: o Teorema de Moran em conjunto com outras ferramentas como Diagrama de Espalhamento de Moran e Indicadores Locais de Associação Espacial (Local Indicators of Spatial Association ou LISA).

Concluiu-se que as medidas de lockdown realmente causaram diminuição da mobilidade nas regiões analisadas. Também observou-se que a menor adoção do isolamento social no Brasil refletiu em grandes aglomerações regionais durante a pandemia. Associou-se tal observação com o fato de que o Brasil foi o país com maior número de casos diários de números de óbitos, no período analisado.

4. Medidas de distanciamento social para o enfrentamento da COVID-19 no Brasil: caracterização e análise epidemiológica por estado

Este artigo tem como objetivo caracterizar as medidas de distanciamento social implementadas pelas unidades federativas brasileiras, para isso descreve-se o tipo de medida adotada e o momento de sua implementação.

Foram utilizados dados retirados de sites oficiais das Secretarias de Governo e do Diário Oficial de cada UF, e por meio da plataforma Brasil.io, que possui dados oficiais dos boletins e informes das Secretarias Estaduais de Saúde de todas UF, obteve-se informações referentes aos casos de Covid-19 e óbitos causados pela doença. Considerou-se seis categorias para as medidas tomadas, são elas: suspensão de eventos, suspensão de aulas, quarentena para grupos de risco, paralisação econômica (parcial ou plena), restrição de transporte e quarentena para a população.

Para análise dos dados não foram feitos pré-processamentos muito complexos, apenas a reagrupação das medidas de distanciamento social. Uma interface gráfica foi usada para uma melhor visualização de dados e tabelas.

Os resultados obtidos foram que a maioria das UF implementaram as medidas antes do primeiro óbito, que poucas foram as medidas realizadas à nível federal, e que em 74% das UF o tempo entre a implementação da

primeira medida e a paralisação econômica foi de aproximadamente uma semana.

5. **Sobre a eficiência de barreiras sanitárias restritivas para conter o avanço da COVID-19: Uma modelagem matemática simples**

Este artigo tem como objetivo estudar os impactos das barreiras sanitárias restritivas impostas em municípios brasileiros, as quais se mostraram insuficientes em sua maioria, além de buscar situações que apresentarem efeitos expressivos. Para atingir tal objetivo foi utilizado um modelo de biologia matemática simples implementado em Fortran, e cálculos probabilísticos.

Para estimar o nível de distanciamento social retirou-se dados do Google Covid-19 Community Mobility Reports e do Mapa brasileiro da Covid-19. Devido aos diferentes tipos de deslocamento também utilizou-se também dados sobre a mobilidade pendular retirados do IBGE .

Os resultados obtidos foram que nesses municípios as barreiras sanitárias não foram eficientes para conter o avanço da doença, mas causaram um atraso considerável no pico de prevalência epidêmica, “achatando a curva”. Observou-se também que em lugares onde as medidas foram mais restritas e severas houve reflexos na taxa de contaminação e avanço da doença.

6. **Análise do efeito das medidas de contenção à propagação da COVID-19 em Belo Horizonte (23/03 a 29/03)**

Este relatório apresenta resultados de estudos que avaliaram o efeito das primeiras medidas de isolamento social, com dados do período de 23 até 29 de março, buscou-se estimar os valores de taxa de transmissão do vírus e assim simular cenários futuros para o número de infectados e a demanda de internação hospitalar.

Para alcançar tais objetivos foi utilizado um modelo SEIR visando obter novas estimativas da taxa de transmissão, técnicas Bayesiana para estimar a taxa de reprodução (evolução do número de infectados). Os dados utilizados foram fornecidos pelo município de Belo Horizonte.

As análises sugeriram que embora as medidas preventivas tomadas pela população tiveram um efeito positivo, apenas depois de medidas tomadas pelos órgãos públicos, a taxa de transmissão irá apresentar mudanças mais concretas.

III - MATERIAIS E MÉTODOS

1. Base de Dados

Os dados referentes aos casos de Covid-19 foram coletados no Open DataSUS e os referentes à taxa de Isolamento Social foram obtidos no portal do estado de São Paulo.

Tais dados foram pré-processados e filtrados por meio de um script em Node.js. Eliminou-se dados implausíveis, como por exemplo, dados de casos com datas anteriores ao início da pandemia, dados de pessoas com mais de 300 anos, entre outros. Selecionou-se também apenas dados das cidades em que a taxa de isolamento social estava disponível no portal do estado de São Paulo. Por fim, mesclou-se os dados referentes ao isolamento social com os dados referentes aos casos semanais e obteve-se um banco de dados organizado no formato JSON. Iniciou-se com um arquivo JSON de 748,6 MiB contendo dados sobre 924 172 casos e finalizou-se com um arquivo JSON de 2,2 MiB contendo 3888 entradas referentes às semanas em cada uma das cidades analisadas.

Em seguida criou-se um DataFrame com os dados. Tal conjunto de dados possui 19 atributos (4 categóricos e 15 numéricos) e 3888 entradas relativas a cada semana em cada cidade analisada.

Os atributos são:

- averageAge (média de idade dos infectados na semana);
- mostInfectedSex (sexo mais infectado na semana | F - Feminino, M - Masculino, E - igual);
- weekDayWithMostCases (dia da semana com mais infecções | 0-6 Domingo-Sábado);
- mostCommonTest (tipo de teste mais utilizado na semana)
- daysSinceFirstCase (dias desde a primeira infecção);
- percentageInfected (porcentagem da população infectada na semana)
- percentageInfectedLastWeek (porcentagem da população infectada na semana passada)
- newCases (número de casos novos na semana)
- lastWeekCases (número de casos novos da semana passada)
- averageDistancingRate (média de distanciamento social na semana)
- averageDistancingRate1 (média de distanciamento social na semana anterior)
- averageDistancingRate2 (média de distanciamento social duas semanas atrás)
- averageDistancingRate3 (média de distanciamento social três semanas atrás)
- distancingRateDifferrence1 (averageDistancingRate1 - averageDistancingRate2)
- distancingRateDifferrence2 ((averageDistancingRate1 - averageDistancingRate2) + (averageDistancingRate1 - averageDistancingRate3)) / 2

- growthFactor (newCases / lastWeekCases)
- growthFactorIncreased (growthFactor >= 1.1) (Atributo alvo)
- city (cidade dos casos)
- weekOfYear (número da semana do ano)

2. Exploração e Pré – processamento

Inicialmente realizou-se uma análise visual, visando identificar os pontos interessantes da base de dados. Utilizou-se de histogramas com a média das idades dos infectados, gráficos de barras identificando o dia da semana com mais infecções, e também o sexo mais infectado na semana, além de boxplots para o fator de crescimento, e para a diferença entre média de distanciamento social na semana anterior e a média de distanciamento social de duas semanas atrás.

Visando uma análise mais profunda que correlaciona o número de casos novos semanais e a taxa de isolamento social semanal confeccionou-se um gráfico que mostra ambas informações sobrepostas. Devido à não imediatez da eficácia do isolamento social utilizou-se de deslocamentos, chamados de shifts, para que fosse possível analisar os novos casos semanais em relação à taxa de isolamento social de 1, 2, 3 ou 4 semanas atrás.

Durante o pré-processamento de dados do DataFrame, os valores ausentes foram substituídos pela média ou moda da coluna, padronizou-se os dados, e as variáveis categóricas foram transformadas em numéricas através do uso do OneHot Encoder.

Posteriormente selecionou-se as variáveis plausíveis à serem mantidas para o algoritmo de classificação.

Logo, excluiu-se as variáveis que poderiam vazar informações diretas sobre o alvo, como por exemplo: growthFactor (basicamente o que se quer analisar), newCases, mostCommonTest, averageDistancingRate (variáveis indisponíveis levando em conta que os dados presentes seriam apenas os de semanas anteriores à semana da análise), entre outras.

Remove-se também dados prévios à semana 10, pois os dados referentes ao isolamento social só estão disponíveis a partir do dia 05/03/2020 (semana 10).

3. Seleção de Atributos

Para uma análise mais eficiente são selecionadas as 10 variáveis que possuíram a maior correlação com o alvo. A matriz de correlação foi calculada por meio do coeficiente de Pearson e elencou as 10 variáveis como sendo: *percentageInfectedLastWeek, weekOfYear, daysSinceFirstCase, averageDistancingRate3, averageDistancingRate1, averageDistancingRate2, lastWeekCases, distancingRateDifference1, x0_Arujá* e *distancingRateDifference2.*

Posteriormente realizou-se uma redução de dimensionalidade por meio do método PCA, mantendo a variância explicada maior que 90%. Por fim, visando uma execução mais rápida dos algoritmos de classificação normalizou-se os dados.

4. Modelos utilizados

Foram analisados 5 classificadores: Perceptron, KNN (com K=5), Multi-Layer Perceptron (15,), SVM Polinomial Grau 3 e Árvore de Decisão Critério Gini.

Para a escolha do melhor classificador, que conseguisse definir se o fator de crescimento (*growthFactor*) seria inferior à 1,1 com base nos dados fornecidos, analisou-se primariamente a métrica de precisão. Tal métrica foi escolhida, pois julgou-se mais importante a eficiência na classificação correta de semanas que possuem fator de crescimento inferior à 1,1, às que possuem fator maior ou igual a 1,1. As análises realizadas foram feitas através de histogramas das precisões dos classificadores com e sem o uso do KFold com K = 10. Em ambas situações o modelo Multi-Layer Perceptron se mostrou melhor, com precisão de aproximadamente 65%.

Analisou-se também outras métricas, como: acurácia, sensibilidade e F1 Score. Com exceção da sensibilidade, que teve como melhor classificador o SVM Polinomial, o melhor classificador para todas estas métricas também foi o Multi-Layer Perceptron.

Por fim, analisou-se a área debaixo da curva ROC, e observou um melhor desempenho do classificador Multi-Layer Perceptron, que obteve uma área de aproximadamente 0,7.

IV - EXPERIMENTOS E DISCUSSÕES

Visualização de dados:

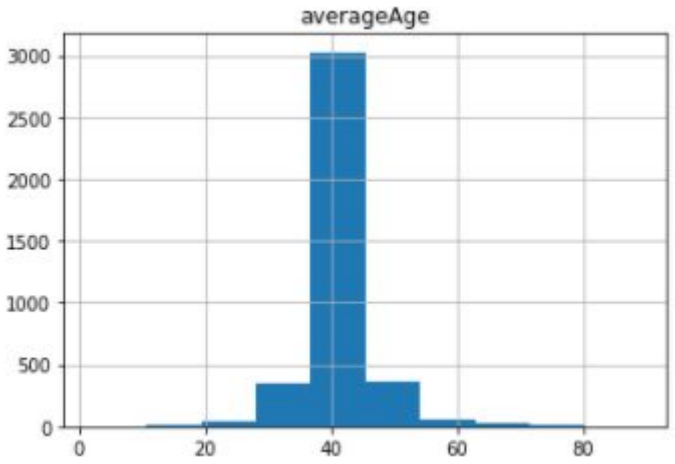

Figura 1 - Histograma: distribuição da idade dos infectados
Obs. mostra uma distribuição normal, algo esperado

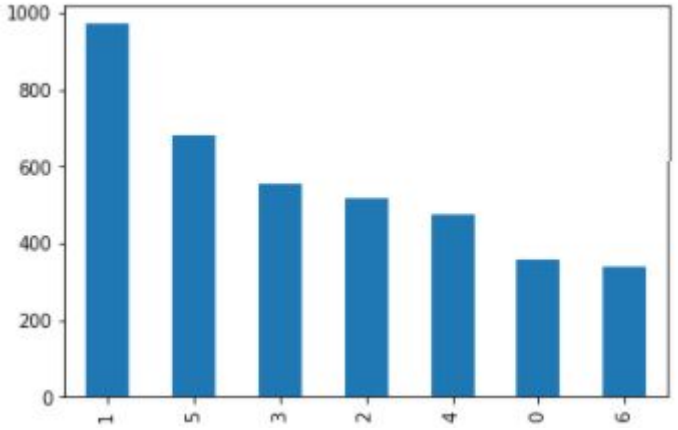

Figura 2 - Gráfico de barras: dias da semana com mais casos (0-6, domingo-segunda)

Obs. Segunda-feira é o dia da semana em que mais casos foram confirmados, provavelmente, devido ao fato de que durante o fim de semana o número de testes feitos é menor e prefere-se a realização dos testes no primeiro dia útil.

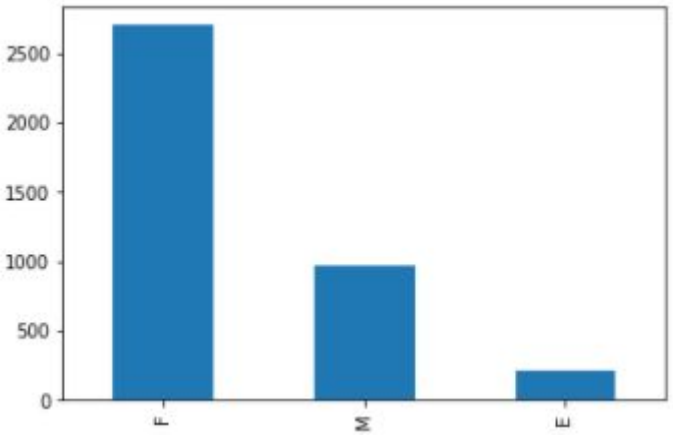

Figura 3 - Gráfico de barras: sexo mais infectado

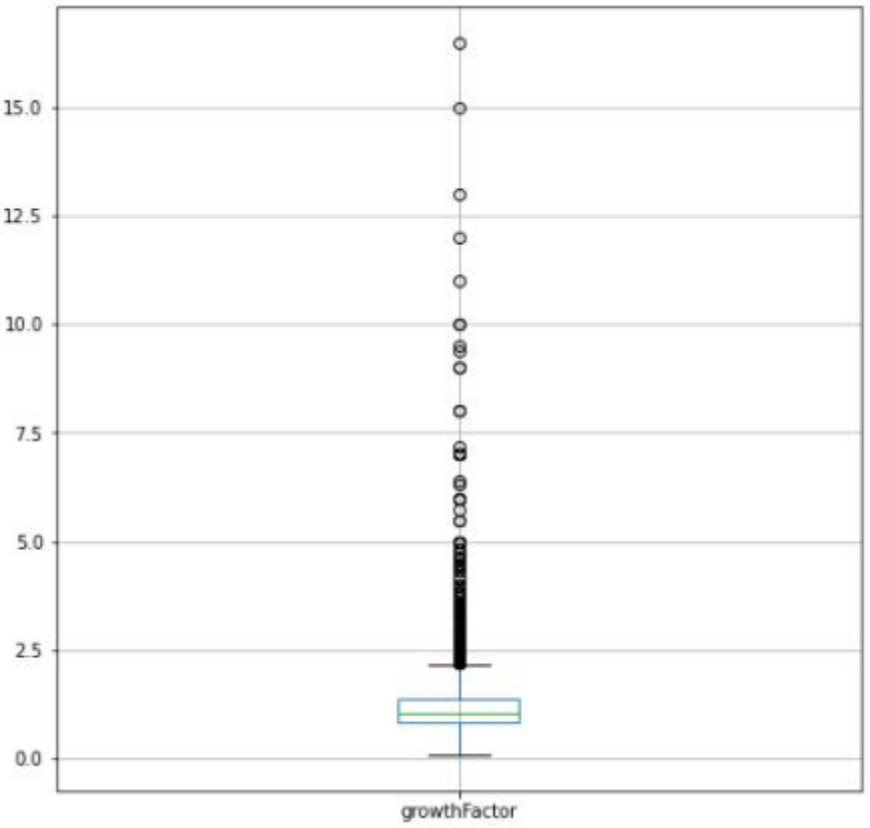

Figura 4 - Boxplot: Fator de crescimento

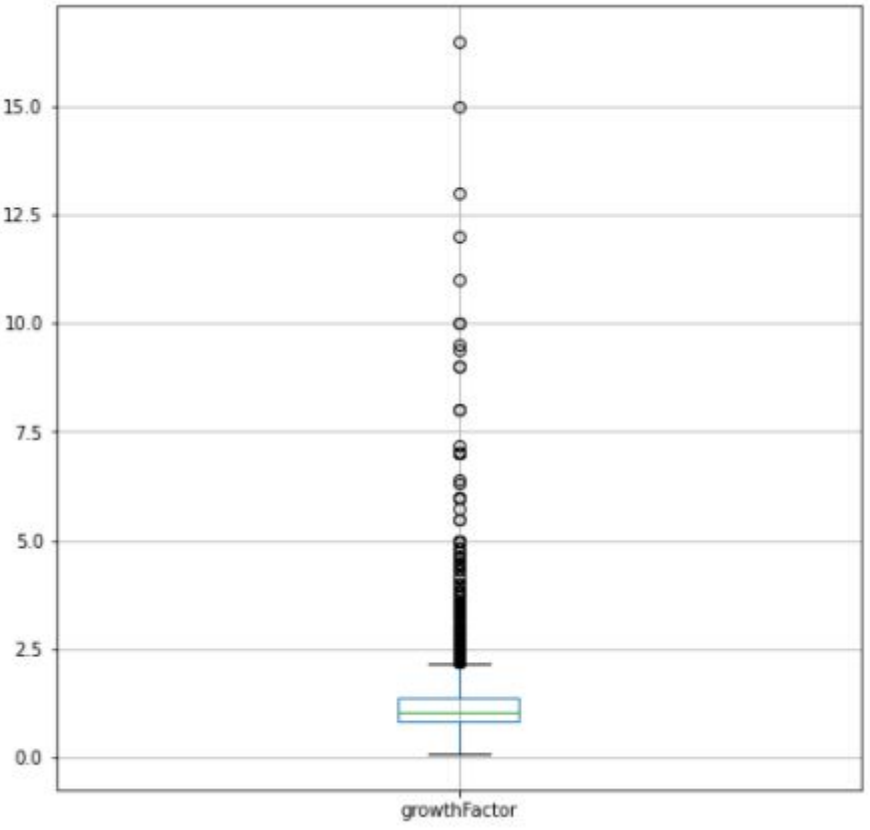

Figura 5 - Boxplot: média do distanciamento social semanal

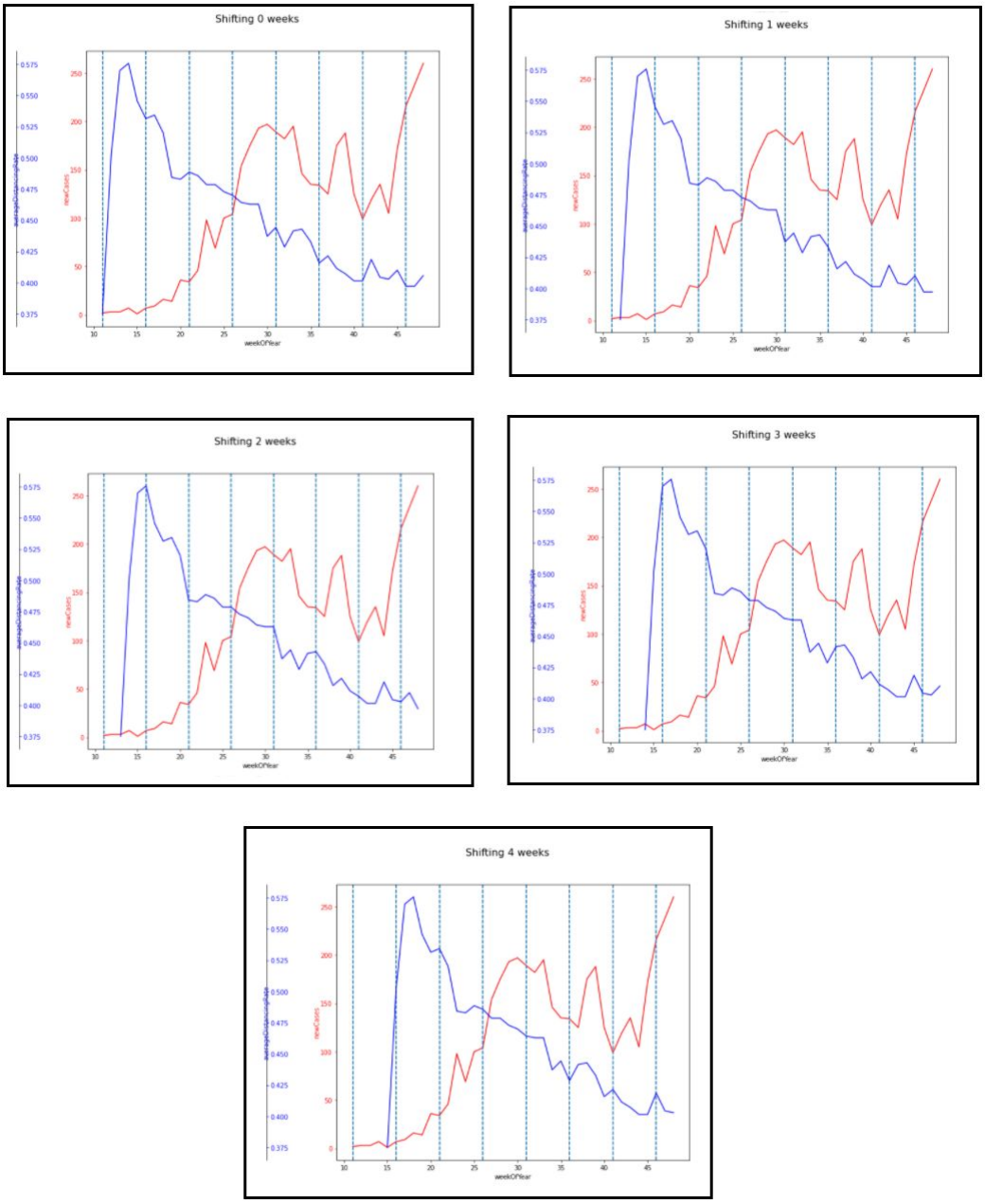

Figura 6 - Taxa de isolamento semana (azul) x Número de novos casos (vermelho)
(Exemplo: cidade de São Carlos)

Resultados:

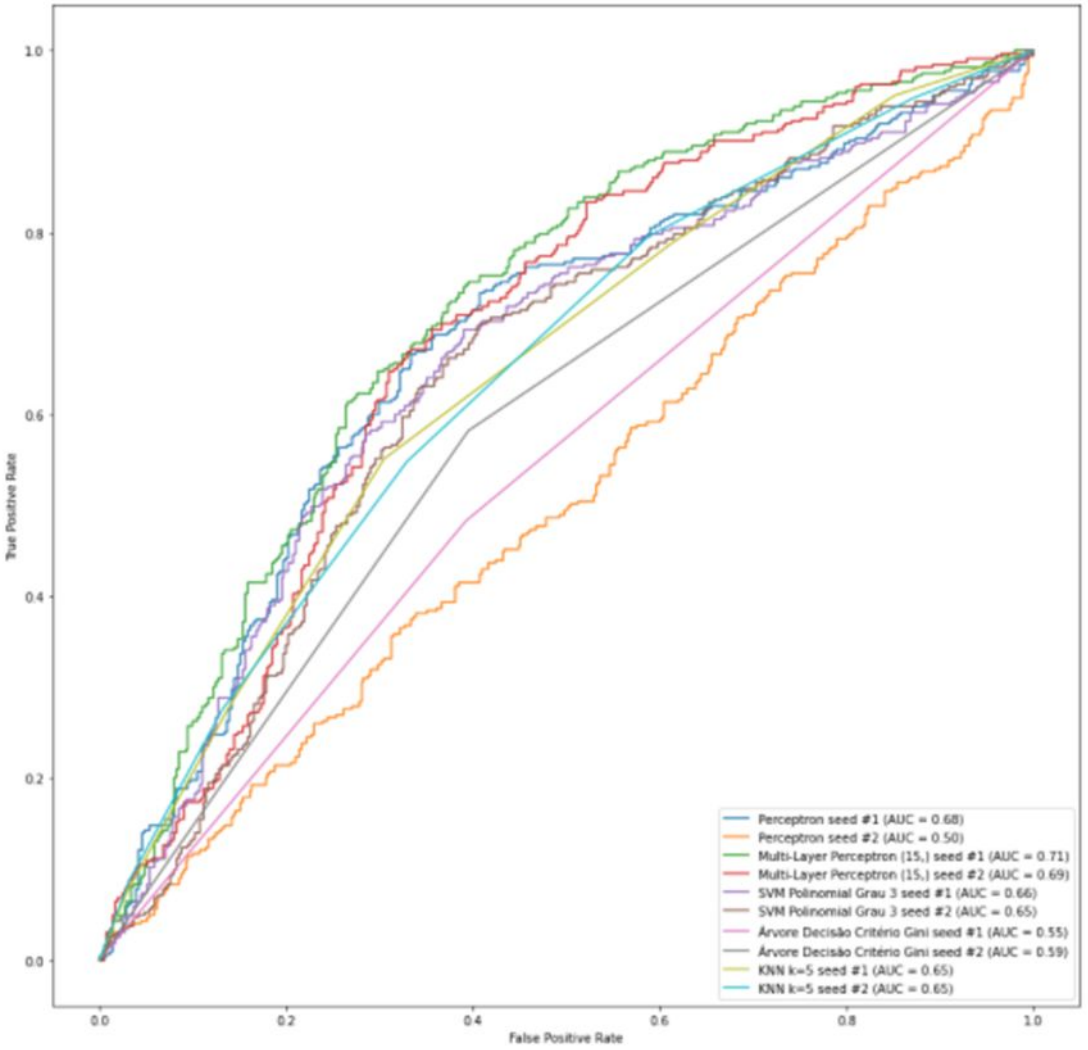

Figura 7 - Curva ROC de todos classificadores

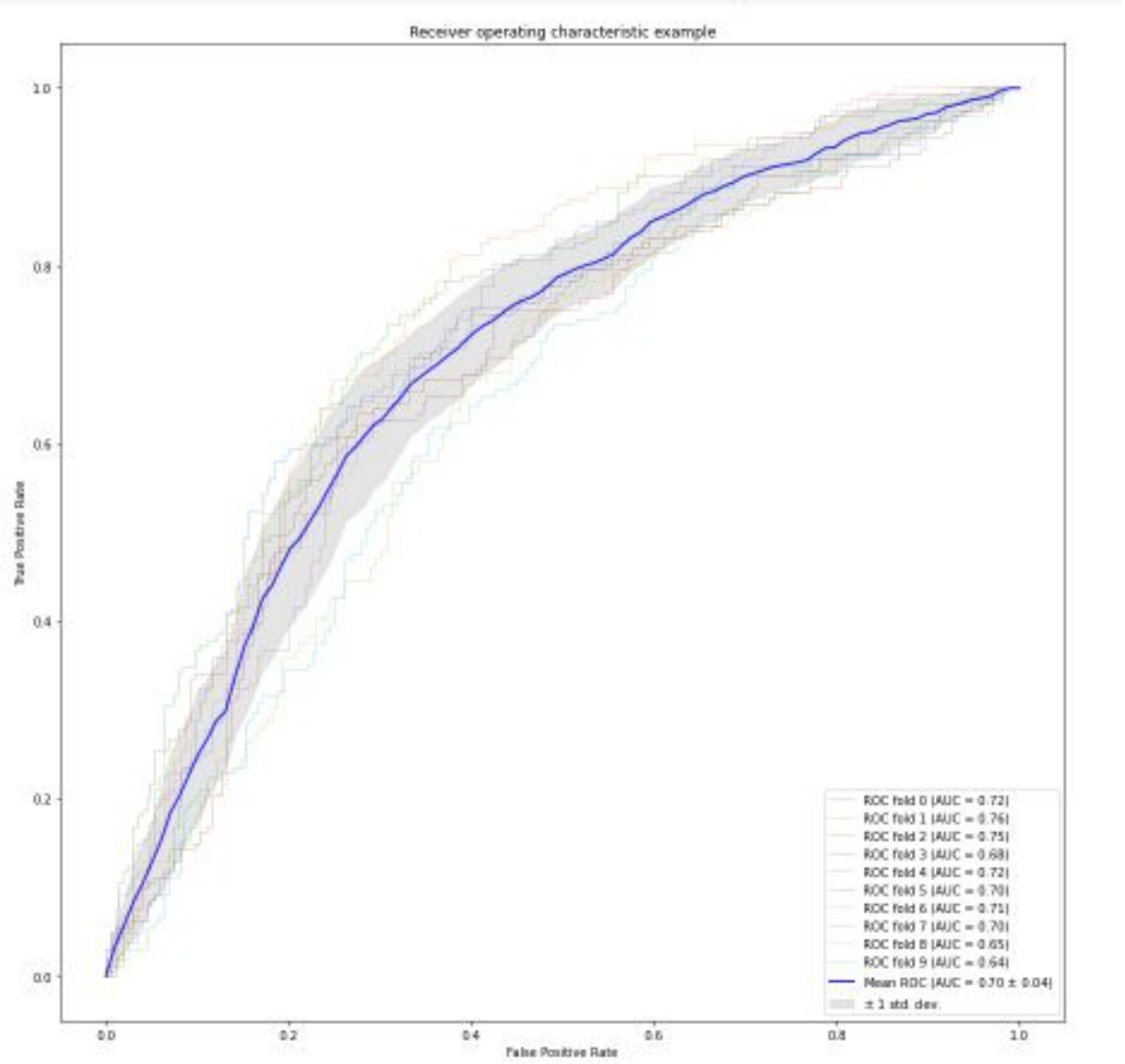

Figura 8 - Curva ROC do Multi-Layer Perceptron

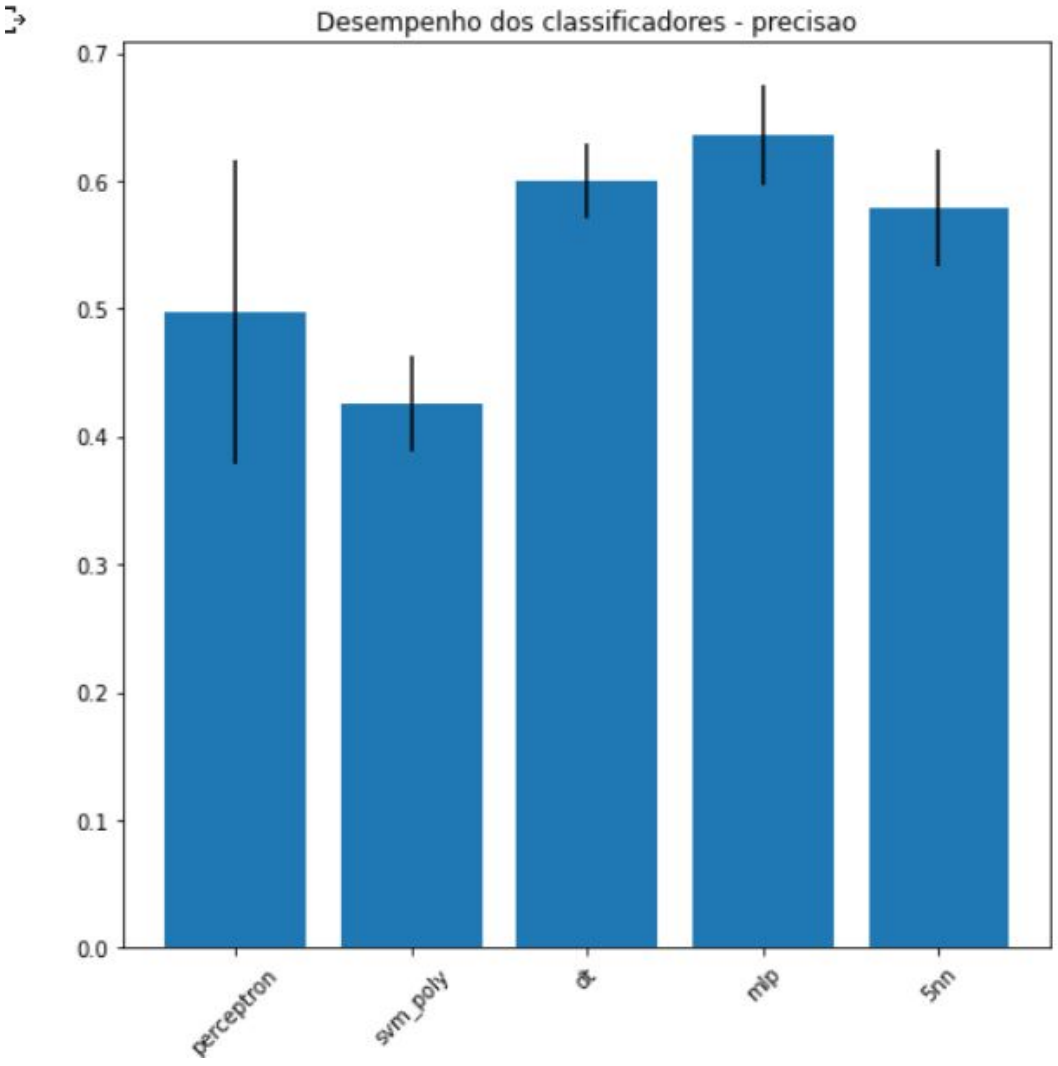

Figura 9 - Histograma da precisão dos classificadores analisados

V - CONCLUSÕES

Após extensa análise e discussão, é possível observar que o distanciamento social, bem como o tempo desde o início da pandemia são fatores importantes para prever o crescimento ou não do número de casos novos da Covid-19. Observa-se também que o melhor classificador dentre os avaliados é o Multi-Layer Perceptron. É importante pontuar também, que mesmo que uma área embaixo da curva ROC de 0,7 não seja o melhor cenário para um algoritmo de classificação, a fidelidade dos dados pode ter causado tal problema, e espera-se que com dados mais fidedignos seja possível obter resultados melhores.

REFERÊNCIAS

Brasil. Mistério da Saúde. Secretaria de Vigilância em Saúde. (2020). Boletim epidemiológico especial - Doença pelo Coronavírus COVID-19. *Boletim Epidemiológico*.

Briz-Redón, Á., & Serrano-Aroca, Á. (2020). The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. *Progress in Physical Geography*. https://doi.org/10.1177/0309133320946302

da Silva, L. L. S., Lima, A. F. R., Polli, D. A., Razia, P. F. S., Pavão, L. F. A., Cavalcanti, M. A. F. de H., & Toscano, C. M. (2020). Medidas de distanciamento social para o enfrentamento da COVID-19 no Brasil: caracterização e análise epidemiológica por estado. Cadernos de Saúde Pública. https://doi.org/10.1590/0102-311x00185020

de Oliveira, G. L. A., de Lima, L. C., Silva, I., Ribeiro-Dantas, M. da C., Monteiro, K. H., & Endo, P. T. (2020). Medidas de distanciamento social e mobilidade na América do Sul durante a pandemia por COVID-19: Condições necessárias e suficientes ? Arvix (Pré-Print):2006.04985v1.

Xu, R., Rahmandad, H., Gupta, M., DiGennaro, C., Ghaffarzadegan, N., Amini, H., & Jalali, M. (2020). The Modest Impact of Weather and Air Pollution on COVID-19 Transmission. *SSRN Electronic Journal*. https://doi.org/10.1101/2020.05.05.20092627