Modelo de Previsão com Regressão Polinomial Para Casos de COVID-19 na Cidade de Tianguá-CE Através dos Classificadores *Support Vector Machine* e *Random Forest*

Roney Nogueira de Sousa [1] **Ronieri Nogueira de Sousa** [2],
Rhyan Ximenes de Brito[1], **Janaide Nogueira de Sousa Ximenes** [2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Av. Tabelião Luiz Nogueira de Lima S/N – Tianguá – CE – Brazil

[2]Faculdade IEducare (FIED) – Rua Conselheiro João Lourenço,
406 - CEP 62320-000 – Tianguá – CE – Brasil

{rxbrito,nogueiraroney453,nsronieri,nogueirajanaide}@gmail.com

Abstract. *COVID-19 is one of the biggest public health problems faced in Brazil and in the world today. For this research, data from the Brasil.io platform were used. Thus, polynomial regression algorithms were used to predict the cases of COVID-19, based on training and testing based on data extracted from the daily bulletins that are provided by the Municipal Health Department. The results obtained were satisfactory as they were close to those observed in reality. Thus, the classifier Random Forest (RF) obtained the best results with an 83.70% average rate of the coefficient of determination compared to the 54.50% obtained by Support Vector Machine (SVM).*

Resumo. *A COVID-19 é um dos maiores problemas de saúde pública enfrentados no Brasil e no mundo atualmente. Para essa pesquisa utilizou-se de dados disponibilizados pela plataforma Brasil.io. Dessa forma usou-se de algoritmos com regressão polinomial para predizer os casos de COVID-19, com base no treinamento e teste a partir dos dados extraidos dos boletins diários que são fornecidos pela Secretaria de Saúde do Município. Os resultados obtidos foram satisfatórios visto que foram próximos aos observados na realidade. Assim, o classificador Random Forest (RF) obteve os melhores resultados com 83,70% de taxa média do coeficiente de determinação frente aos 54,50% obtidos pelo Support Vector Machine (SVM).*

1. Introdução

Em dezembro de 2019, foi notificada à Organização Mundial da Saúde (OMS) um surto de pneumonia na cidade de Wuhan na China. Os resultados da investigação laboratorial das amostras de lavado bronquioalveolar identificaram um novo coronavírus como responsável pelo surto [Chaves and Bellei 2020].

O novo coronavírus comprovadamente é transmitido de humano para humano, através de contato próximo, por gotículas respiratórias eliminadas na tosse ou espirro, inalação de aerossóis, objetos contaminados e pela via fecal-oral[Aquino et al. 2021].

A partir da utilização de técnicas computacionais que auxiliem na previsão de aumento de casos da doença. Esse trabalho tem como principal objetivo realizar um estudo

utilizando os classificadores *Support Vector Machine* (SVM) e *Random Forest* (RF), buscando obter um modelo de previsão de casos da *COVID-19* na cidade de Tianguá-CE. O estudo foi realizado a partir de um banco de dados público disponibilizado através do *link*: https://brasil.io/dataset/covid19/files/.

Este trabalho está organizado da seguinte forma: Seção 2 apresenta a fundamentação teórica sobre a *COVID-19* e os classificadores *Support Vector Machine* e *Random Forest*, a Seção 3 apresenta a metodologia utilizada, e por fim a Seção 4 apresenta as considerações finais.

2. Fundamentação Teórica

Esta seção apresenta uma revisão teórica sobre os assuntos abordados nesse trabalho. Com a Subseção 2.1 abordando definições e características da COVID-19. A Subseção 2.2 e Subseção 2.3 com abordagem teórica sobre os classificadores *Support Vector Machin*e e *Random Forest* com suas suas particularidades.

2.1. *COVID-19*

Os coronavírus pertencem à subfamília *Coronavirinaena* família *Coronaviridae* da ordem *Nidovirales*, e esta subfamília inclui quatro gêneros: *alfacoronavírus*, *betacoronavírus*, *gamacoronavírus*, e *deltacoronavírus* [Chaves and Bellei 2020]. Este vírus possui estrutura membranosa de espinhos proteicos e penetra nas células através dos receptores celulares da Enzima Conversora de Angiotensina 2 [de Campos Tuñas et al. 2020].

Segundo [Estevão 2020] estima-se que aproximadamente 80% dos doentes desenvolvam doença leve, 14% doença grave e 5% doença crítica. Os doentes com doença grave geralmente apresentam sinais e sintomas de pneumonia viral e podem evoluir para situações de Síndrome de Dificuldade Respiratória Aguda, insuficiência cardíaca aguda, lesão renal aguda, sépsis ou choque.

2.2. *Support Vector Machine* (SVM)

O SVM é um algoritmo de aprendizado de máquina supervisionado que identifica padrões, podendo este ser aplicado para classificação e regressão. Fundamenta-se no conceito de planos de decisão que definem os limites de decisão, construindo hiperplanos [Moraes and Machado 2009].

O hiperplano tem a função de alocar a maioria dos pontos da mesma categoria de um mesmo lado. A subclasse de amostras de dados mais próximos do hiperplano são os vetores de suporte [Moraes and Machado 2009].

2.3. *Random Forest* (RF)

O RF é um algoritmo de aprendizagem de máquina baseado em árvores de decisão. Elas são treinadas isoladamente na tentativa de encontrar um modelo para resolver o mesmo problema diminuindo a variância. Mostra-se muito eficiente quando se busca analisar um grande volume de dados [de Alvarenga Júnior 2018]. Este modelo é usualmente utilizado não apenas para classificação, mas também para regressão, estudo de importância e seleção de variáveis, e detecção de *outliers* [de Alvarenga Júnior 2018].

3. Metodologia

Para o emprego dos algoritmos utilizou-se para implementação a linguagem Python e um banco de dados público composto por 369 amostras disponibilizadas através do link: https://brasil.io/dataset/covid19/files/, além de dados adicionais, como períodos de *lock-down* e fechamento de comércio. Os resultados das regressões polinomiais foram extraídos com base nos classificadores SVM e RF com a técnica de validação cruzada *k-fold*, com k=10 *folds* e a normalização (*z-score*). Salienta-se que os dados antes do início do processo de *data mining* (DM) foram pré-processados excluindo-se amostras de dados faltosos ou que não possuiam relação com a classificação, como por exemplo código do município. Para as saídas esperadas utilizou-se as quantidades de casos diários.

4. Resultados e Discussões

Os resultados obtidos e analisados tiveram como base o coeficiente de determinação e o erro médio absoluto adquiridos nos treinamentos e testes realizados.

Conforme observado na Tabela 1 o melhor caso com o classificador *Random Forest* atingiu um coeficiente de determinação de 93,40% e um erro médio absoluto 7,09, enquanto o pior caso atingiu um coeficiente de determinação de 78,54% com o erro médio absoluto de 15,70. A média do coeficiente de determinação foi de 83,70% e do erro médio absoluto de 12,47.

Tabela 1. Resultados Random Forest

Folds	Coeficiente de Determinação(%)	Erro Médio Absoluto	Situação
1	88,00	12,10	
2	80,05	14,03	
3	80,20	14,84	
4	82,21	13,04	
5	84,27	12,41	
6	78,54	15,70	Pior Caso
7	81,17	13,50	
8	90,30	8,90	
9	93,40	7,09	Melhor Caso
10	81,64	12,88	
total	83,70	12,47	Caso Médio

Já para os resultados com o SVM a Tabela 2 mostra que para o melhor caso atingiu para o coeficiente de determinação 81,21% e um erro médio absoluto 12,88, enquanto que o pior caso atingiu 25,69% para o coeficiente de determinação e 50,42 para o erro médio absoluto. A média do coeficiente de determinação foi de 54,50% e do erro médio absoluto foi de 32,78.

Se comparados os resultados para o caso médio entre os dois classificadores (RF, SVM) será observado que com o RF houve um ganho de 29,20% para o coeficiente de determinação, para o erro médio absoluto o SVM teve um aumento de 20,31. Porém se analisado os melhores e piores casos entre os classificadores (RF, SVM) será percebido que para o melhor caso, o RF obteve um ganho de 12,19% para o coeficiente de determinação e o SVM um aumento de 5,79 para o erro médio absoluto. Para o pior caso

Tabela 2. Resultados Support Vector Machine

Folds	Coeficiente de Determinação(%)	Erro Médio Absoluto	Situação
1	70,25	18,45	
2	72,65	22,69	
3	59,68	30,41	
4	25,69	50,42	Pior Caso
5	34,37	42,37	
6	48,69	34,63	
7	45,23	38,42	
8	27,98	48,45	
9	64,74	29,16	
10	81,21	12,88	Melhor Caso
total	54,50	32,78	Caso Médio

observou um ganho de 52,85% coeficiente de determinação com o RF e um aumento de 34,72 do erro médio absoluto com o SVM.

5. Considerações Finais e Trabalhos Futuros

O artigo relatou a utilização de classificadores para a previsão do aumento de casos da doença COVID-19 na cidade de Tianguá-CE com base em um banco de dados composto por boletins diários a qual relatam o comportamento do vírus na cidade.

Foi constatado que o classificador *Random Forest* apresentou a a melhor média atingindo uma média do coeficiente de determinação de 83,70% e o erro médio absoluto 12,47, enquanto o classificador SVM obteve o pior resultado com média do coeficiente de determinação de 54,50% e o erro médio absoluto 32,78, apresentando também uma grande variância em seus resultados, o que faz com que esse classificador não seja indicado para o problema proposto. Como trabalhos futuros sugere-se a utilização de redes neurais comparando os resultados obtidos considerando as métricas usadas no RF e SVM.

Referências

Aquino, R., Alves, J., and Carvalho, J. (2021). Transmissão vertical do novo coronavírus. *Monumenta-Revista Científica Multidisciplinar*, 2(1):29–36.

Chaves, T. S. and Bellei, N. (2020). O novo coronavirus: uma reflexao sobre a saude unica (one health) e a importancia da medicina de viagem na emergencia de novos patogenos. *Revista de Medicina*, 99(1):i–i.

de Alvarenga Júnior, W. J. (2018). Métodos de otimização hiperparamétrica: um estudo comparativo utilizando árvores de decisão e florestas aleatórias na classificação binária.

de Campos Tuñas, I. T., da Silva, E. T., Santiago, S. B. S., Maia, K. D., and Silva-Júnior, G. O. (2020). Doença pelo coronavírus 2019 (covid-19): Uma abordagem preventiva para odontologia. *Revista Brasileira de Odontologia*, 77:1–7.

Estevão, A. (2020). Covid-19. *Acta Radiológica Portuguesa*, 32(1):5–6.

Moraes, R. M. and Machado, L. S. (2009). Gaussian naive bayes for online training assessment in virtual reality-based simulators. *Mathware & Soft Computing*, 16(2):123–132.