Feature engineering of an exogenous variable for a viral disease prediction model

Jonatas Silva do Espirito Santo, Gecynalda Soares da Silva Gomes, Rodrigo Barbosa de Cerqueira, Camila Braz Soares

The spread of diseases resulting from human viral infections can occur at a speed above the capacity of health agents to register notifications. As a result, using methods to predict the behavior of epidemic curves is necessary to support the implementation and direction of public actions to combat these diseases. An indicator of the history of searches on symptoms, diagnoses and treatments of Covid-19 that the population made on the internet can be used as a predictor of the number of infections. In this work, we studied the alignment of the curve of new cases of COVID-19 in Bahia and the curves of the Relative Search Volumes (RSV), the search index for terms in Google Trends. The alignment between the COVID-19 case curve and the RSV curve of some terms was calculated using the Dynamic Time Warping (DTW) algorithm. The term “covid test” was chosen as it was one of the expressions that the RSV curve is more aligned with that of new cases of COVID-19. Also, the correlation between the curve of confirmed cases of coronavirus and the RSV curves lagged in until ten epidemiological weeks was assessed, using Pearson's correlation coefficient. Results point to a greater correlation (r=0.85) between the curves of new COVID-19 cases and that of RSV lagged by 3 weeks. Since the alignment of the curves presents different behavior in specific periods, the series was divided into intervals referring to the waves of COVID-19 in Bahia. The first wave presents greater detachment between the series and the highest correlation occurs only for lags of 7 weeks (r=0.95). The second wave showed the highest correlation, with the RSV curve lagging by one week (r=0.87). The third and fourth waves have a higher series correlation with a lag of 3 and 2 weeks, respectively, with coefficients of 0.98 and 0.96. Time series of Google Trends search indexes for the term “covid test” is a potential predictor of COVID-19 case curves, anticipating the number of SARS-CoV- infections by up to three weeks in Bahia.