Better use of data. The project falls into this category because its objective is to use an AI for data correlation analysis.

George Augusto Moreira Czelusniak, Jessica Vaz Scheffer, Julia Eunice Fernandes, Lucas de Jesus Gonçalves, Rafael Felipe Tasaka de Melo

May 2020

1 Abstract

Our team has developed, in NASA Space Apps COVID-19 Challenge, an artificial intelligence (A.I.) that performs a correlation between human development data and COVID -19 contamination. Thus, it is possible to base public policies and prevention methods on these results, which are based on data from the United Nations, Datasus and the Health Secretary of each municipality. Our proposal can be replicated worldwide so as to predict, anywhere, the spread of the virus. In this way, it can become the main tool for the eradication of the pandemic, in addition to being a model for the idealization of similar tools that will combat future pathologies.

2 Introduction

Since the beginning of the pandemic, several studies have started proposing correlations between population density and the spread of COVID-19, however there are several other indices to be taken into account, such as human development indicators, which can be considered for the construction of scenarios that allow a forecast of the next points of dissemination.

Since the beginning of the pandemic, several studies have started proposing correlations between population density and the spread of COVID-19, however there are several other indices to be taken into account, such as human development indicators, which can be considered for the construction of scenarios that allow a forecast of the next points of dissemination.

Some authors have addressed human development issues and their relation with COVID-19 (as a COVID-19 spreading factor), such as Mogi et. al (2020) [4], who worried to relate socioeconomic inequalities and the spread of the pandemic, as many indicators imply that economically disadvantaged cities tend to have fewer confirmed cases of infection than economically favored ones. In other words, the association that richer people have more cases is due to the fact that they have a greater number of commuting outside cities (such as commuting to work), facilitating propagation, while those with lower socioeconomic status make movements in a smaller radius.

The need to seek a correlation between aspects of human development and the dissemination of COVID-19 arose from the ambiguity found among some case studies, such as the one that took place in some US municipalities by Mukherji (2020)[5]. In his study, it was identified that cities with higher GINI indexes; social inequality quantifier ranging from 0 to 1, the closer to 0 the lower the inequality; also had a greater number of confirmed Coronavirus cases. While a case from South Africa, a country with a high GINI coefficient (0.63), made by Stiegler and Bouchard (2020)[6], showed a great control of the infected curve.

Applied studies in Rio de Janeiro, based on socioeconomic inequality in order to measure the impact of interventions to reduce the spread of COVID-19, carried out by Klôh et. al (2020)[7], used Artificial Intelligence (AI) to identify the populations most exposed to pathology. It was concluded that people from locations with greater social inequality are more exposed to the disease.

Considering the bibliographic references and the objective of the challenge passed by NASA Space Apps, the Hotspots team developed an AI to find the correlations between human development indicators and the spread of the Coronavirus.

3 Methodology

As suggested in the first Nasa Space Apps webinar on guidelines for the hackathon, a brainstorming methodology was carried out in order to find the central problem in which it would be worked. After gathering suggestions on topics with similar themes, the most relevant themes were selected by election based on a debate and the problem more aligned with the proposed Human Factors Challenge was identified.

Having defined the problem, identifying hot spots, namely, predicting areas of greater dissemination of COVID-19, Crazy-8, a methodology for creating quick solutions, was executed. In this methodology, team members had to, in eight minutes, focus on possible solutions for the proposed problem. After the solution's development, each member presented their ideas, allowing them to be grouped according to the theme similarity. After a discussion, three topics were ranked by influence and impact on the topic: a) comparison of Human Development Index in areas where the virus spreads, b) increasing the quantity of tests or differ tests methods to enhance the database and c) political measures against COVID-19. Finally, after some arguments, it was decided to focus on the first theme, given the possibility of finding quantitative references and the short deadline.

Then, based on a literature survey, a database of brazilian cities was retrieved from the United Nations Development Program (UNDP)[3], a United Nations (UN), where human development indexes were obtained, Datasus[2], where was found the Gini coefficient of towns; and from health secretariats linked to the brazilian ministry of health, collected the gross population and confirmed cases[1]. This survey aimed to identify possible patterns in order to predict the areas that may be affected in the future, creating a model with the potential to be replicated in other countries. With the research by city, the following data was retrieved: Gini coefficient, HDI (Human Development Index), responsible for comparing the degree of economic development and quality of life between countries, number of cases and deaths confirmed by COVID-19, it was possible to give start the development of an AI.

The model of AI used was a multiple linear regression that consists in an analysis of many variables. The databases consist of 9 attributes being them: name of cities, COVID-19 confirmed cases, death by COVID-19, population, Gini coefficient, general MHDI, income MHDI, longevity MHDI and schooling MHDI; of 4179 cities. To test the algorithm, attributes like “name of cities” and “death by COVID-19” were eliminated.

The database was randomly divided in a way that 75% of its total was designated to train the A.I. and the remaining 25% was used for tests. To analyse the results, the Mean Absolute Error (MAE) parameter was employed.

In order to understand the influence of towns with a lower number of confirmed cases in problem modeling, the same database was compiled 9 fold, the first with no restrictions and the other 8 with given different restrictions. On the second, it was only analysed data from cities which have more than 100 confirmed cases; on the third, using towns with more than 500 confirmed cases. Interactions fourth to eighth had a regular increase on the minimum number of 500 per interaction, being the ninth and last one with 5000 of the same parameter. For each interaction, a different learning model was generated.

4 Results and discussion

The results from MAE are shown on Table 1

From various tests in the algorithm, it’s noticed that the greater the quantity

Table 1: Comparison between data and Mean Squared Error

MEAN SQUARED ERROR IN EACH TYPE OF DATABASE	
Database	**MAE**
All data	163.364
More than 500 cases	1479.606
More than 1000 cases	6069.681
More than 1500 cases	5467.466
More than 2000 cases	4856.801
More than 2500 cases	7162.523
More than 5000 cases	24653.380

with small number of cases are left off the analysis, the higher MAE becomes, thus it is cogitable that those with more confirmed cases are more susceptible to total errors.

Small towns commonly have an expressively low amount of confirmed cases. From that, it may be troublesome to use them for a general analysis, once they can invalidate the analysis with data that does not match reality due the late or yet to come pandemic outbreak. Using a greater amount of higher rate of confirmed cases cities implies a high error to smaller cities, which can be used to predict the pandemic growth rate on small towns, once then are a consequence of larger cities that shares the same index.

References

[1] Brazilian epidemiological report covid-19.

[2] Brazilian gini index of per capita income.

[3] Brazilian mhdi ranking 2010.

[4] Ryohei Mogi;Gento Kato;Susumu Annaka. Socioeconomic inequality and covid-19 prevalence across municipalities in catalonia. 2020.

[5] Nivedita Mukherji. The social and economic factors underlying the impact of covid-19 cases and deaths in us counties. *medRxiv*, 2020.

[6] Nancy Stiegler and Jean-Pierre Bouchard. South-africa: challenges and successes of the covid-19 lockdown. *revue psychiatrique*, 2020.

[7] Mariza Ferro Eric Araújo Cristiano Barros de Melo José Roberto Pinho de Andrade Lima Ernesto Rademaker Martins Vinícius Prata Klôh, Gabrieli Dutra Silva. The virus and socioeconomic inequality: An agent-based model to simulate and assess

the impact of interventions to reduce the spread of covid-19 in rio de janeiro, brazil. *revue psychiatrique*, 2020.