COVID-19 in Mexico: Identifying health and socioeconomic variables for death and hospitalization prediction.

CAPP 30254: Machine Learning for Public Policy

Roberto Barroso Luque* Oscar E. Noriega Villarreal†
Jesica María Ramírez Toscano‡ Luz Stephanie Ramos Gómez§

DRAFT: June 9, 2020

Abstract

The current pandemic caused by the SARS-CoV-2 (COVID-19) virus represents an unprecedented challenge to the global economy. In order to provide data-driven policy advice to individual countries analysis and modeling should be done on national-level data. With under-funded healthcare systems, institutional-weakness, and large informal economic sectors, Mexico faces a gigantic challenge to contain the virus. We developed classification algorithms to predict death and hospitalization among individuals who tested positive for the virus in Mexican soil using a mix of health and socioeconomic variables at the individual and municipality level. Using ten-fold cross validation and synthetic minority over sampling linear support vectors, logistic regression, decision trees and random forest models were trained to predict patient outcomes (survival vs death and hospitalization vs home recovery). In both classification problems, recall was used as the main evaluation metric in order to minimize false negatives. The best models achieved accuracy scores of more than seventy percent and recall scores of more than eighty percent. Feature importance for the optimum models were then analyzed. Age and comorbidities such as diabetes, immuno-suppresion and obesity gave strong

*barrosoluquer@uchicago.edu
†onoriega@uchicago.edu
‡jramireztoscano@uchicago.edu
§stephanieramos@uchicago.edu

predictive power to both COVID-19 death and hospitalization predictions. Surprisingly, health resource indicators such as number of doctors, nurses and hospital beds per 100,000 residents as well as municipality poverty level featured among the top predictors for hospitalization but not for death prediction. A relative risk profile was then constructed from the classification results. Our work shows that, despite inaccuracies in governmental data, machine learning can provide insights into important risk factors that may be used to forecast severity of COVID-19 outcomes at the individual level while informing policy at the state level.

1 Introduction

On January 30th, the World Health Organization (WHO) declared the SARS-CoV-2 outbreak a Public Health Emergency of International Concern, yet despite the institution's warnings, political leaders have failed to act in a timely fashion.
As of June 2020, about 7 million people have been infected, and more than 400,000 have died from the COVID-19 pandemic. Until a vaccine is developed, which experts predict will take more than a year, the most efficient policy to hinder viral transmission is mass social-isolation. Unfortunately, while social isolation measures stem the spread of the virus they also have a devastating economic effect.

Millions of jobs have been lost, children have stopped going to school, entire economies are in standby. Countries, such as Canada and the US have provided emergency funds to help middle- and low-income households through social protection mechanisms. In contrast, developing countries' fiscal response has been timid at best and non-existent at worse. Economies such as Mexico and India, where at least half of the labor population works in the informal sector (meaning no access to social protection and/or financial institutions), face additional obstacles.

In Mexico, the Central Bank anticipates a GDP decline of up to 8.8%[1] in a worst-case-scenario, while private institutions such as Credit Suisse expect a drop of 9.6%[2] in GDP. Given its limited resources, it is of imperative importance that the Mexican government

[1] *Quarterly Report, January-March 2020*- Banco de Mexico
[2] *Mexico's GDP will have its worst decline since 1932: Credit Suisse* - Forbes Mexico, 04/30/2020

prioritize the most vulnerable areas and allocates resources accordingly. However, due to a lack of research and analysis, it is far from clear which states and municipalities are most at risk.

The purpose of this project is to fill this analysis gap to identify vulnerable communities and speed up the governmental response. To do so, we will rely on a Machine Learning approach for:

- Identifying the best socioeconomic and health predictors for COVID-19 deaths and hospitalizations among positive cases.
- Gaining insights into the most at risk populations and locations based on key predictors to inform resource allocation.

Using ten-fold cross validation and Synthetic Minority Over-sampling Technique (SMOTE), we trained four different Machine Learning models: Linear Support Vectors, Logistic Regression, Decision Trees and Random Forest in order to predict COVID-19 patient outcomes (survival vs death and hospitalization vs home recovery).

Using recall as our main evaluation metric, in order to minimize false negatives, the best models achieved recall scores of more than eighty percent and accuracy scores of more than seventy percent.
Our feature importance analysis revealed that age and comorbidities such as diabetes, immuno-suppresion and obesity gave strong predictive power to both COVID-19 death and hospitalization predictions. Surprisingly, municipal-level data in the form of health resource indicators (such as number of doctors, nurses and hospital beds per 100,000 residents) and poverty level featured among the top predictors for hospitalization but not for death prediction.
A relative risk profile was then constructed from the classification results. [INDEX RESULTS].

The paper is structured as follows: Section 2 provides a brief overview of the COVID-19 situation in Mexico, Section 3 describes the data and data sources used, Section 4 establishes the Machine Learning problem we're dealing, as well as briefly describing the models

used for our predictions, Section 5 shows our main results and Section 7 the conclusions of our analysis.

2 COVID-19 in Mexico

Following a now-familiar trend, Mexican authorities, fell prey to excessive hubris as the coronavirus started to spread around the world.

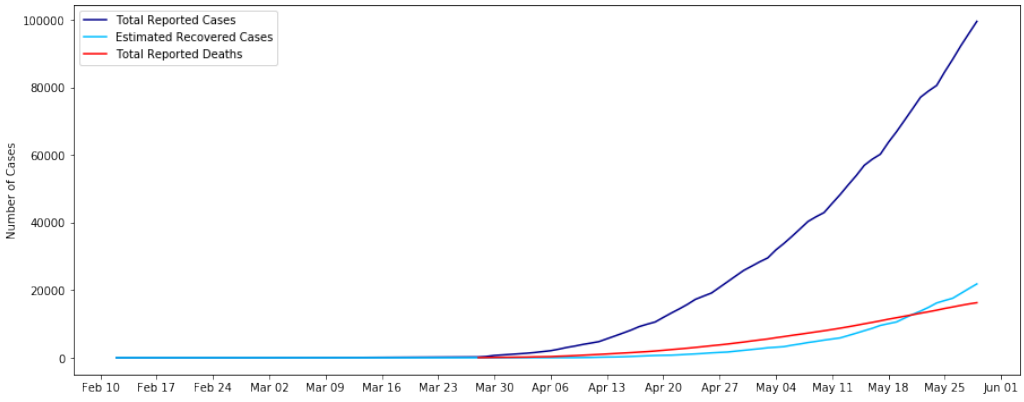
In late January, official communication from the Ministry of Health downplayed the risks presented to the Mexican population by the quickly spreading virus[3]. By the end of February, Mexico reported its first official Coronavirus cases one in Mexico City and one in the northern state of Sinaloa. Despite the alarming pace of global transmission, and the mounting death toll in countries like Italy and Spain, by late March the Mexican government continued with its cavalier attitude[4].

As of June 2020, Mexico has over 120,000 official cases and more than 14,000 deaths from the virus. Whether the disturbing state of the pandemic currently unfolding in Mexico is a direct consequence of official government policies or reckless rhetoric from political leaders is a question for academic debate and political punditry[5]. Nonetheless, what is abundantly clear from official data is that Mexico has not managed to control the epidemic or 'flatten the curve' (Figure 1).

Official cases, hospitalizations, and deaths as a result of the novel coronavirus continue to rise across the country. Most alarmingly, the country has an average death rate of more than 11% according to official counts which are substantially larger when compared to similar countries such as Peru (2.8%), Colombia (3%), and Brazil (5%). Mexico's excessive official death rate could be the cause of low testing capacity, as of June 2020 the country's testing rate falls at a meager 2600 tests per million residents, Brazil which is one of the

[3] *New virus doesn't represent a danger for Mexico.* José Pablo Espíndola. Reporte Indigo.
[4] *AMLO's feeble response to COVID-19 in Mexico.* Vanda Felbab Brown. Brookings.
[5] *Mexican President López Obrador draws doctors' ire.* David Agren. The Lancet.

Figure 1: Total COVID-19 reported cases in Mexico



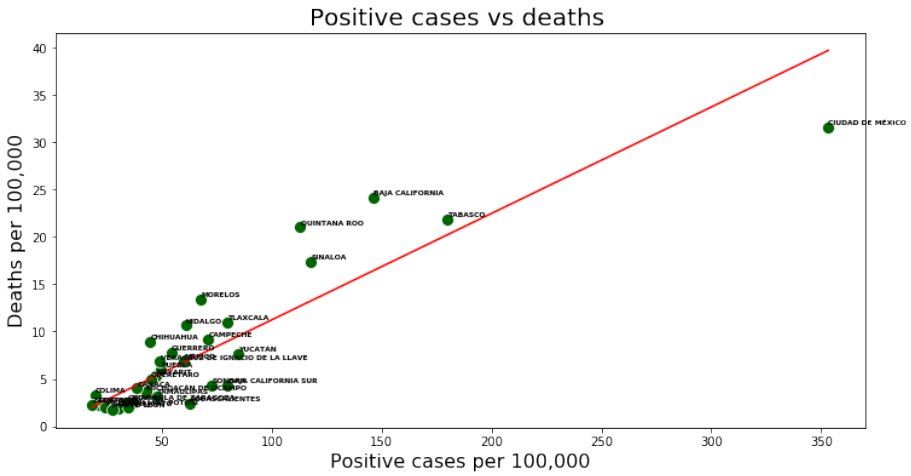
Source: Health Ministry data

worst-hit countries has a test rate of about double that of Mexico[6]. Yet until the testing capacity increases the true gravity of the situation will remain unknown.

The severity of the pandemic has not been uniformly felt across the country with the northern states of Baja California and Sinaloa and southern Quintana Roo being particularly hard hit (Figure 2). Furthermore, as has been the case in most countries, specific segments of the population have been at much higher risk of developing severe outcomes of the disease, with older cohorts and people with comorbidities at much higher risk of dying or being hospitalized as a result of the virus (Figure 3).

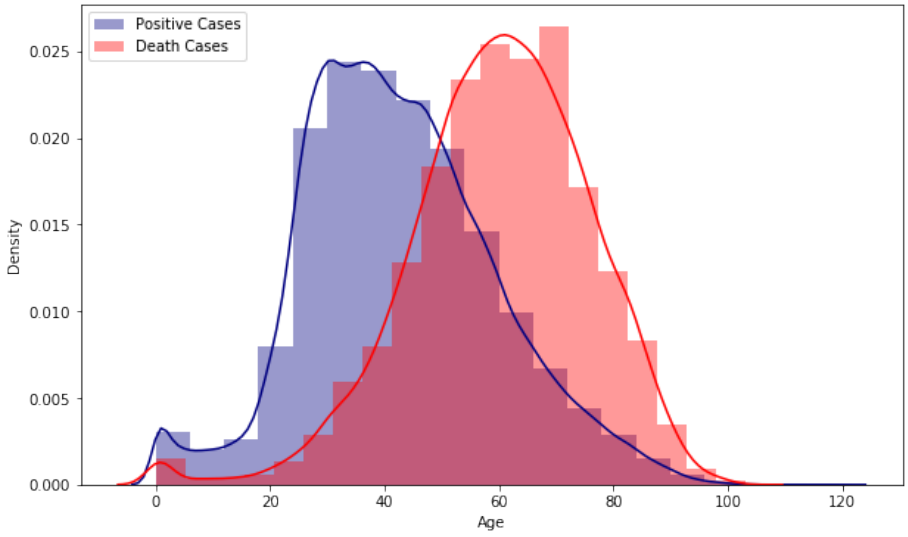
[6] See Statista data

Figure 2: COVID-19 positive cases vs. deaths by state in Mexico



Source: Health Ministry data

Figure 3: Age distribution for reported COVID-19 cases in Mexico



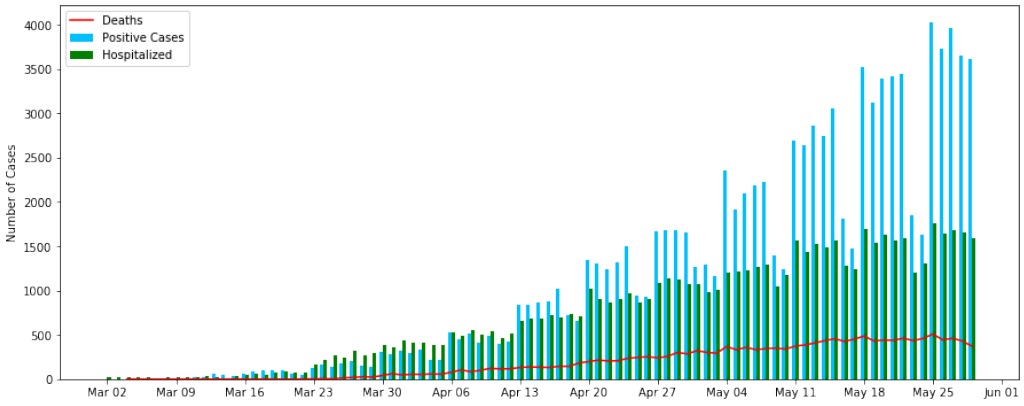
Source: Health Ministry data

3 Data

The data used from this analysis was collected from a variety of sources. While our unit of observation is the individual, we also care about relevant characteristics of the municipality where the individual lives, particularly important here are the health infrastructure available.

For the individual data, we use the daily reports from Mexico's Federal Health Ministry (SSA)[7]. This dataset[8] provides information on all individual COVID-19 positive cases reported in Mexico and contains our outcomes of interest: whether an individual infected with COVID-19 dies or gets hospitalized (Figure 4), with age, state, and municipality specified for each individual. The dataset also contains data on the specific comorbidites that the individual presents at the time of testing positive for COVID-19, such as diabetes, obesity, inmuno-supression, asthma, among others.

Figure 4: COVID-19 Daily New Cases, Hospitalizations and Deaths in Mexico



Source: Health Ministry data

Since not only individual characteristics define our outcomes of interest, we also need to account for features at the municipality level. Mainly, it's important to include data on

[7] *Secretaría de Salud*

[8] *datos.gob.mx COVID-19 Open Data*

Health Infrastructure and available personnel at the municipality level, since we have data on the municipality that each individual lives in. We used another SSA database[9] that contains information on physical resources (beds, clinics, operating rooms)and health personnel (doctors, nurses, technicians, and other personnel).
It's important to notice that this dataset only accounts for federally-funded health facilities; while we're missing data on state, municipal and privately funded hospitals, according to the National Council of Social Development Policy Evaluation (CONEVAL), more than 90% of the reported COVID-19 cases are treated in federally-funded facilities[10].

Additionally, we used the National Population Council (CONAPO) 2020 projections for municipality population and CONEVAL's data on municipality poverty rates, available for all 2,464 municipalities in Mexico.

To build one DataFrame for our analysis, we first merged all the municipality-level data (Health resources, population and poverty) and then merged this municiaplity-level data with the COVID-19 individual-level data, in order to have a dataset where the observation unit is the individual, but contains information on the individual's municipality.

4 Machine Learning Problem

> *'AI applications present opportunities for the future of healthcare and can be harnessed at this time, as clinicians take on the complexities of responding to COVID-19'*
>
> *Jiang et al. (2020)*

Given that resources are highly limited, the Mexican government must prioritize the most vulnerable communities in a timely and efficient manner. In this context, vulnerable communities can be places where the unemployment or poverty levels could substantially increase, or where there is low/limited access to health resources or where the population has a higher risk of dying if infected with COVID-19.

[9] *datos.gob.mx COVID-19 Health Resources 2018*
[10] *Poverty and COVID-19 Findings*- CONEVAL

In this report, we define a vulnerable community as a place where the individuals who have COVID-19 are at higher risk of ending up in a severe condition (i.e. hospitalized or dead) compared to other communities.
One way to identify who will be more likely to end up in a severe condition is to predict COVID-related deaths or hospitalizations. In this sense, with Mexican COVID-19 data, we trained several classification models to predict if a COVID-19 patient will die or be hospitalized.

As previously stated, we have a single DataFrame containing data both at individual and municipality level with our features and targets defined as follows:

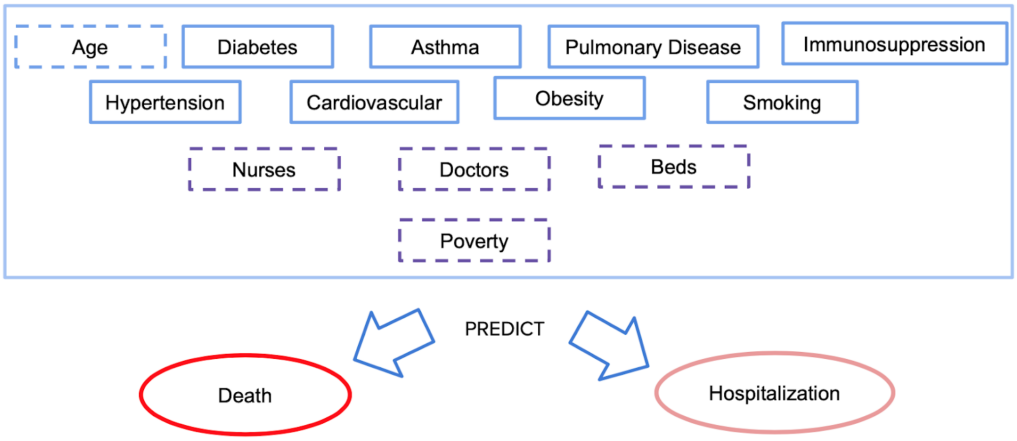
- Feature variables (continuous variables in dashed rectangles, binary ones in solid)
  - At the individual level (in blue)
  - At municipality level (in purple):
    - Nurses, doctors, and hospital beds per 10, 000 people
    - Percentage of people living in poverty in the municipality
- The target variables (in red) to predict for each patient: death and hospitalization (binary variables 1 positive outcome, 0 negative outcome)

To predict whether an individual with COVID-19 ends up hospitalized or dead, we use several classification models such as: logistic regression, support vector machines, decision trees, and random forest. All of these models have been applied for clinical use to predict the likelihood of disease or likelihood of risk of recurrence (See: Shipe et al (2019), Viera et al (2013), Caruana et al. (2015)).

4.1 Models

+SMOTE
SUB3
*index building

Figure 5: Features and target variables



5 Main Results

5.1 Prediction

5.2 Risk Index

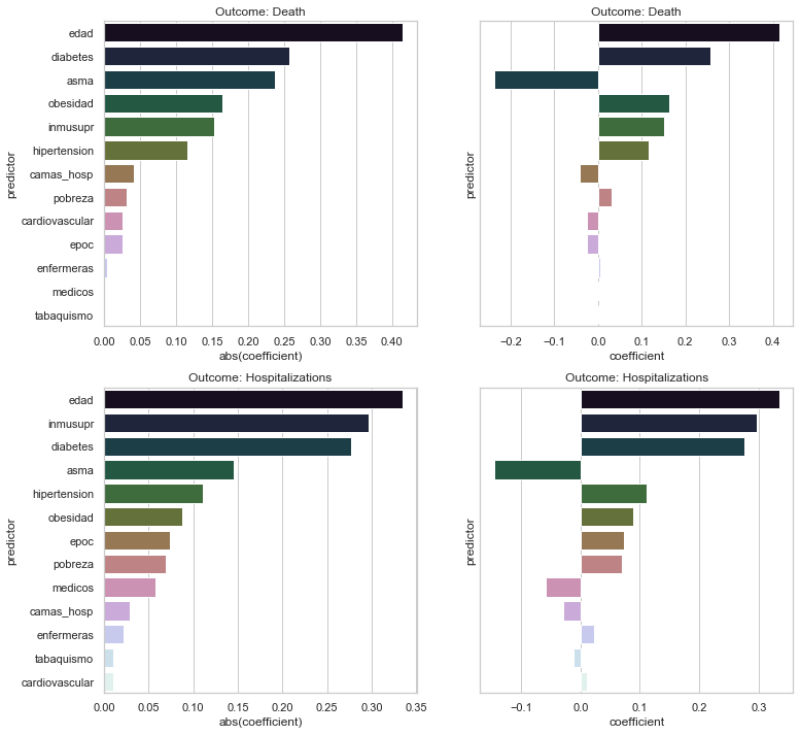
Table 1: Evaluation metrics for COVID-19 related deaths predictions

Prediction outcome: *Death*	Metrics:		
	Recall	*Precision*	*Accuracy*
Linear Support Vector	0.73	0.23	0.71
Logistic Regression	0.71	0.24	0.72
Decision Tree	0.87	0.19	0.58
Balanced Random Forest	0.82	0.21	0.65
Weighted Random Forest	0.77	0.23	0.69
Complement Naive Bayes	0.74	0.15	0.54

Table 2: Evaluation metrics for COVID-19 related hospitalization predictions

Prediction outcome: *Hospitalization*	Metrics:		
	Recall	*Precision*	*Accuracy*
Linear Support Vector	0.82	0.21	0.65
Logistic Regression	0.67	0.56	0.70
Decision Tree	0.88	0.19	0.57
Balanced Random Forest	0.73	0.57	0.71
Weighted Random Forest	0.73	0.57	0.70
Complement Naive Bayes	0.71	0.44	0.59

Figure 6: Feature importance for COVID-19 deaths and hospitalizations prediction using SVM classifier generating synthetic data (SMOTE) per each fold in CV



6 Conclusions and remarks

6.1 Policy Recommendations

6.2 Limitations, Caveats, and Suggestions for Future Work

7 Additional References

- Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease, 11(Suppl 4), S574.

- Vieira, S. Mendonça, L., Farinha, G., Sousa, J. (2013) Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied

Soft Computing.

- Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M. et al. (2015): Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730.

8 Appendix

*additional graphs