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 10, 2020

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

1 Executive Summary

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 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.

2 Background: 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[1]. 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[2].

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[3]. Nonetheless, what is abundantly clear from official data is that Mexico has not managed to control the epidemic or 'flatten the curve' (Supplementary Figure 3).

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 worst-hit countries has a test rate of about double that of Mexico[4]. 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 (Supplementary Figure 4). 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 (Supplementary Figure 5).

Unfortunately, while social isolation measures stem the spread of the virus they also have a devastating economic effect. Countries, such as Canada and the US have provided emergency funds to help middle- and low-income households through social protection mechanisms. In contrast, Mexico's fiscal response has been timid at best and non-existent at

[1] *New virus doesn't represent a danger for Mexico.* José Pablo Espíndola. Reporte Indigo.

[2] *AMLO's feeble response to COVID-19 in Mexico.* Vanda Felbab Brown. Brookings.

[3] *Mexican President López Obrador draws doctors' ire.* David Agren. The Lancet.

[4] See Statista data

worst. With at least half of the labor population in the informal sector (meaning no access to social protection and/or financial institutions) Mexico's economy is at a particular perilous state, where Mexico's Central Bank anticipates a GDP decline of up to 8.8%[5] in a worst-case-scenario, while private institutions such as Credit Suisse expect a drop of 9.6%[6] in GDP. Given its limited resources, it is of imperative importance that the Mexican government 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. To address this analysis gap we analyzed official COVID-19 cases provided by the Ministry of Health and complemented this data with health resources, demographic, and poverty data at the municipality level from several public sources. We trained classification algorithms on health and socioeconomic variables to predict the severity of COVID-19 outcomes and used our result to develop a relative risk profile for all Mexican states. We hope that this work can serve to inform Mexican politicians and policy leaders so that they can provide data-driven solutions to the current crisis. Furthermore, we hope our analysis will prove useful to the Mexican research and scientific communities so that they continue to work on investigating and understanding the nature of the pandemic in our country.

3 Data

The data used from this analysis was collected from several 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). This dataset[7] 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 (Supplementary Figure 6), with age, state, and municipality specified for each individual. The dataset also contains data on the specific

[5] *Quarterly Report, January-March 2020*- Banco de Mexico

[6] *Mexico's GDP will have its worst decline since 1932: Credit Suisse* - Forbes Mexico, 04/30/2020

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

comorbidites that the individual presents at the time of testing positive for COVID-19, such as diabetes, obesity, inmuno-supression, asthma, among others.

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 Health Resources at the municipality level, since we have data on the municipality that each individual lives in. We used another SSA database[8] that contains information on physical resources (beds, clinics, operating rooms)and health personnel (doctors, nurses, technicians, and other personnel). Additionally, we used the National Population Council (CONAPO) 2020 projections for municipality population and the National Council of Social Development Policy Evaluation (CONEVAL) 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 (row) is the individual, but contains information on the individual's municipality as well.

4 Machine Learning Problem

In order to predict whether an individual with COVID-19 ends up hospitalized or dead, we used the following classification models: Logistic Regression, Support Vector Machine, Decision Trees, Random Forest and Naive Bayes. 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)).

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
  - At the individual level: diabetes, asthma, pulmonary disease, immunosuppression, hypertension, cardiovascular condition, obesity, smoking (all binary except for age).

[8] *datos.gob.mx COVID-19 Health Resources 2018*

  - At municipality level: Number of nurses, doctors, and hospital beds per 10,000 people and percentage of people living in poverty in the municipality (all continuous).

- Target variables to predict for each patient: death and hospitalization (binary variables 1 positive outcome, 0 negative outcome)

Our data set turned to be highly imbalanced, deaths account for 11% of the data and hospitalizations account for 33%. Several classification models do not work well with imbalanced data. Decision trees and random forest models aim to minimize the overall error rate, rather than paying special attention to the positive class (the "rare class"). Logistic Regression has also been proved to underestimate the probability of rare events (See: King and Zeng (2001)). To alleviate the problem, we follow two different approaches when applying a model: one is using a sampling technique called SMOTE and the other is using a modified version of the Random Forest algorithm.

For the alternate Random Forest algorithm, we used two modified models that account for imbalanced data: balanced random forest and weighted random forest (See: Chen et al (2004)). The Balanced Random Forest works by drawing a bootstrap sample from the minority class, at each iteration, and randomly draws the same number of cases, with replacement, from the majority class. The weighted random forest places a heavier penalty on misclassifying the minority class (i.e. the minority class is given a larger weight). The weights are used in the tree induction procedure (the splits) and in the terminal nodes (class determined by weighted majority vote).

For the other models (SVM, logistic regression, complement naive Bayes, and decision tree), we applied the Synthetic Minority Over-sampling Technique (SMOTE). Unlike regular over-sampling techniques that repeat the same information, new examples can be synthesized from the existing examples to boost the minority class. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.

For model training, we used a 10 fold cross-validation. Each fold of the training data is split into training and testing sets. For all models, except the modified random forest algorithms where no sampling technique is used, SMOTE is applied to the training set at each fold during cross-validation. This process is applied to SVM, logistic regression, complement naive Bayes, and decision tree as described by figure 4. Then, the trained model is validated with the test set of the training fold. As we try to minimize the false negatives (no death/hospitalization predictions when it actually happened), we picked the best models with the highest average recall score of the test sets (from the training data).

5 Evaluation and Results

Three metrics were used to evaluate the performance of the ML models that answer the following questions:

- Recall: What proportion of actual positive identifications are correctly identified?
- Precision: What proportion of positive identifications are correctly identified?
- Accuracy: What proportion of predictions are correctly identified?

This study seeks to correctly identify as many individuals at risk of being hospitalized or dying due to COVID-19 as possible with the purpose of allocating resources towards those individuals and preventing the disease from turning severe. In other words, we want to minimize false negatives. For this reason, the preferred evaluation metric in this study is the Recall Score.

Table 1 summarizes the results of the machine learning models using both deaths and hospitalizations as target variables. For Deaths, overall recall reached between 71% and 87%, meaning that the false-negative rate of the models is low. On the other hand, precision only reached between 15% and 24%, meaning that the models predict a high number of false positives. Accuracy is between 54% and 71%. Given that recall is the preferred evaluation metric in this context, the models that performed better were the Balanced Random Forest and the Decision Tree. However, the precision score of the decision tree is below the score of all other models. This is a good example of the tradeoff between precision and recall.

For the case of COVID-related hospitalizations, Table 1 shows that the overall recall reached between 71% and 88%. Precision is higher for this target variable, ranging from 0.19% to 0.57%. Accuracy is between 57% and 70%. The models that reached the highest recall scores are the Linear Support Vector and Decision tree. However, these models have the lowest precision scores. Models like the Balanced Random Forest and Weighted

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

	Target: *Death*			Target: *Hospitalization*		
	Recall	*Precision*	*Accuracy*	*Recall*	*Precision*	*Accuracy*
Linear Support Vector	0.73	0.23	0.71	0.82	0.21	0.65
Logistic Regression	0.71	0.24	0.72	0.67	0.56	0.70
Decision Tree	0.87	0.19	0.58	0.88	0.19	0.57
Balanced Random Forest	0.82	0.21	0.65	0.73	0.57	0.71
Weighted Random Forest	0.77	0.23	0.69	0.73	0.57	0.70
Complement Naive Bayes	0.74	0.15	0.54	0.71	0.44	0.59

Random Forest have higher precision but a slightly lower recall score.

The most important features in predicting COVID-related deaths and hospitalizations across models are age and diabetes. Note that these features do not need to be causal to be predictive. Figure 1 shows feature importance for the SVM model and Figure 2 shows feature importance for the Balanced Random Forest. In the case of the SVM model, other important features besides age and diabetes are individual health characteristics like obesity and asthma. In the case of the Balanced Random Forest, features at the municipality level like the number of doctors and poverty have higher importance.

5.1 Risk Index

The predictions made by the machine learning models can be grouped by region and used to construct a relative risk index between Mexican states. We built this index by calculating the percentage of individuals in each state that the models predicted to be at risk of being hospitalized or dying due to COVID-19. Then, we normalized that number to have a mean of zero and a standard deviation of one. A negative score X means that a state is X standard deviations below the meanwhile a positive score Y means that a state is Y

Figure 1: Feature importance for COVID-19 deaths and hospitalizations prediction using SVM classifier

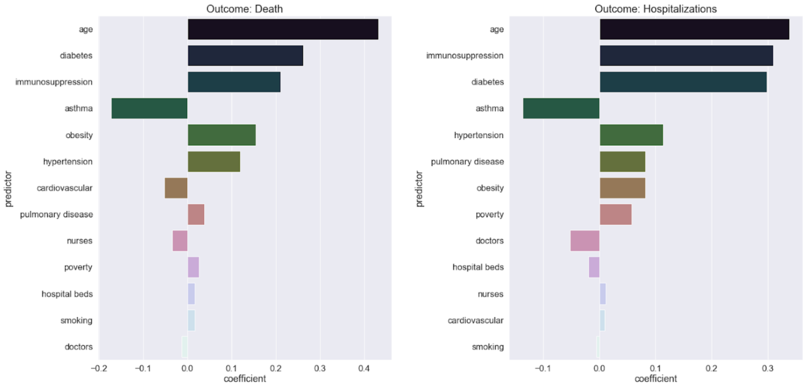
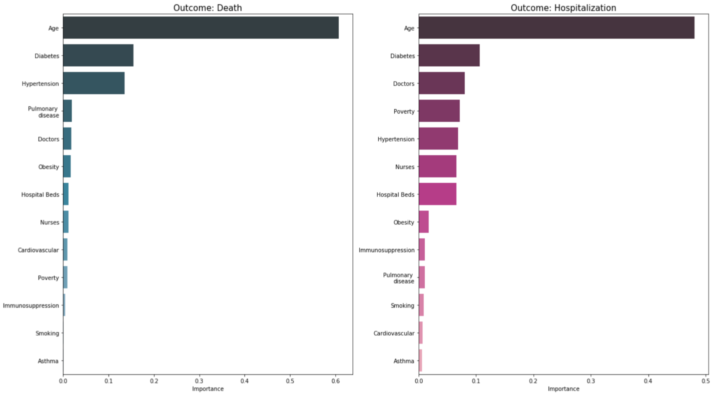


Figure 2: Feature importance for COVID-19 deaths and hospitalizations prediction using Random Forest classifier



standard deviations above the mean[9].

[9] See Supplementary Figure 7 for the results using the Random Forest model and deaths as the target variable and Supplementary Figure 8 for the results using the SVM model and hospitalizations as the target variable.

6 Policy Recommendations

7 Ethics

There are various ethical qualms that arise from our analysis. Due to the COVID-19 reported cases inaccuracies, the results of our analysis are probably biased. Furthermore, due to the lack of access to certain rural regions in the country, we are also concerned that our analysis has ignored or failed to learn statistical relationships from many vulnerable communities that are not recorded (or under-reported) in our data.
To address this we visualized F1-score and recall-score across Mexico to illustrate how these evaluation metrics changed as a function of geography (Supplementary Figure 9). The differences in evaluation metrics that our models achieve as a function of geography are notable for both prediction outcomes (deaths and hospitalizations). Additionally, we are concerned about the possibility that the results of our analysis can be misinterpreted and lead to ill-informed excuses to stigmatize certain groups in the population. For example, an erroneous interpretation of our results would be to conclude that individuals with comorbidities such as diabetes and immunosuppression, which our models indicate are good predictor variables for severe COVID-19 outcome, are in some way more contagious. We implore any reader of our work to not make grandiose conclusions or extrapolate erroneous relationships. The purpose of our work is **NOT** to find causal relationships between health/socioeconomic variables and severe outcomes. Rather, our work helps to identify variables that might be indicative of which individuals are most vulnerable to the virus once infected. This allows policymakers and medical practitioners to prioritize these vulnerable individuals during the triage process and thus offer them the best care possible.

8 Limitations, Caveats, and Suggestions for Future Work

The main caveat to consider in this analysis, as mentioned before, is that the official data set for COVID-19 cases in Mexico that we used is, in all certainty, highly inaccurate. As it has been discussed, Mexico's testing capacity lags far behind many of its peers which means the official case count of infections used in this analysis is much smaller than the real number. Furthermore, it has been suggested that official statistics have underreported the number of deaths as a consequence of the virus[10]. Particularly, we only observe severe cases in the data, since in Mexico (as in many places in the world) only people with harsh symptoms are tested, due to the government reticence for performing mass-testing[11]. Additionally the data set provides very limited information on each individual; having more features could improve the models predictive power. Specifically, direct biochemical measurements such as glucose level, hemoglobin counts, etc.

Regarding the Health Resources data, 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 CONEVAL, more than 90% of the reported COVID-19 cases are treated in federally-funded facilities[12].

As a suggestion for future work, maybe incorporating state-specific data, specifically for health-related statistics can improve the informative nature of the risk index we build.

[10] *Hidden Toll: Mexico Ignores Wave of Coronavirus Deaths in Capital.*- Azam Ahmed. May 2020. The New York Times.

[11] *Mass testing won't happen in Mexico. That's the way the government wants it.*- Matt Rivers. May 2020. CNN.

[12] *Poverty and COVID-19 Findings*- CONEVAL

9 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.

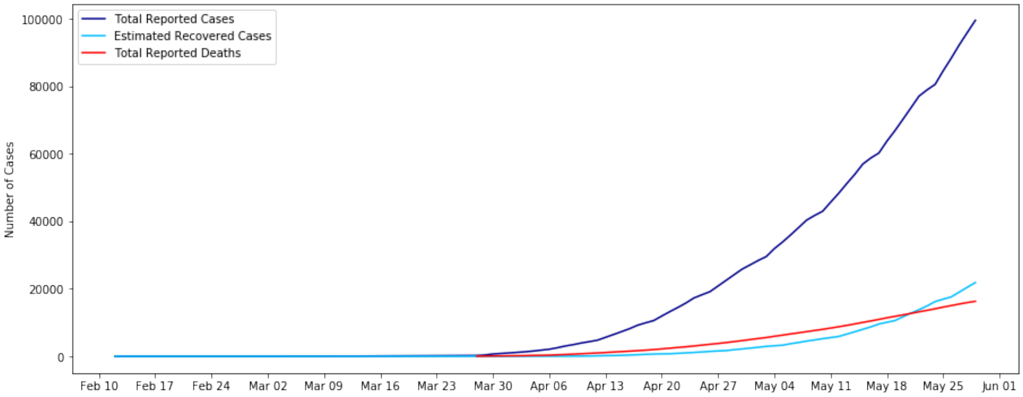
- Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M. et al. (2015): Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730.

- King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political analysis, 9(2), 137-163.

- Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24.

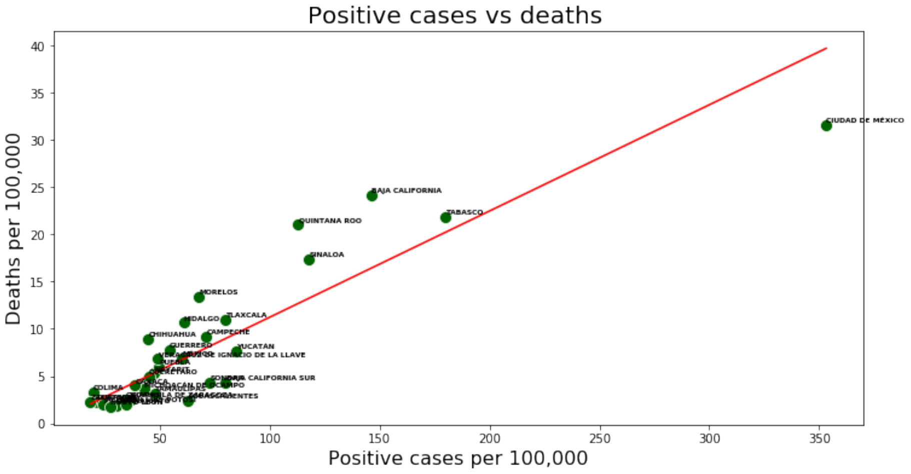
10 Supplementary Figures

Figure 3: Total COVID-19 reported cases in Mexico



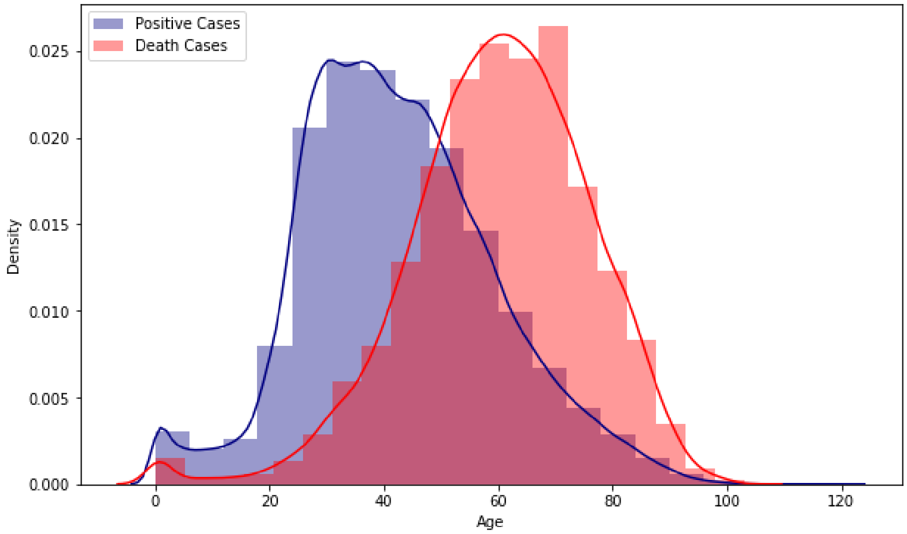
Source: Health Ministry data

Figure 4: COVID-19 positive cases vs. deaths by state in Mexico



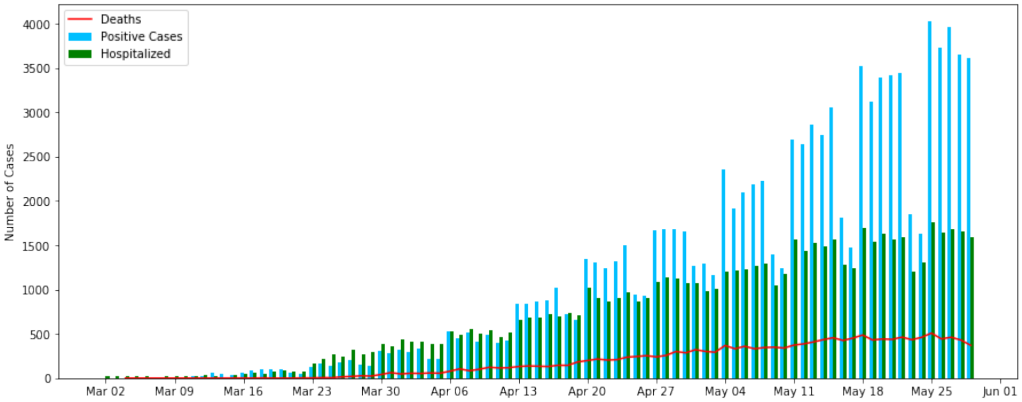
Source: Health Ministry data

Figure 5: Age distribution for reported COVID-19 cases in Mexico



Source: Health Ministry data

Figure 6: COVID-19 Daily New Cases, Hospitalizations and Deaths in Mexico



Source: Health Ministry data

Figure 7: Relative Risk of Deaths by State

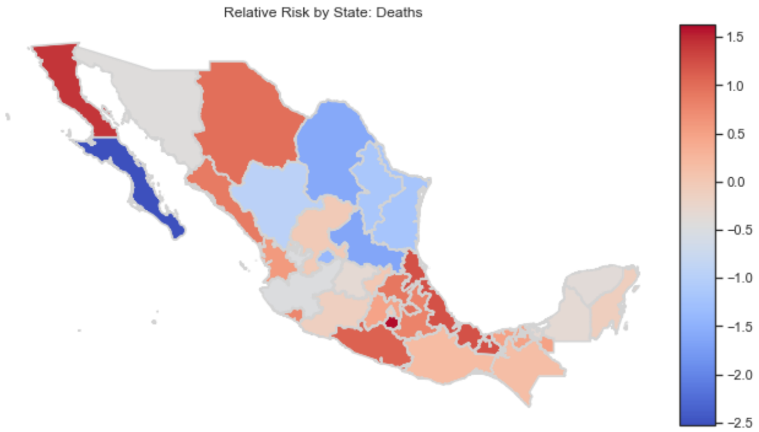


Figure 8: Relative Risk of Hospitalizations by State

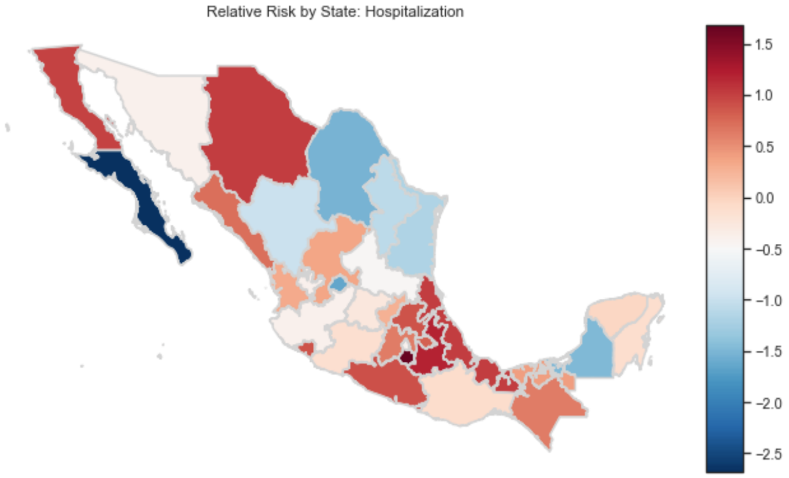
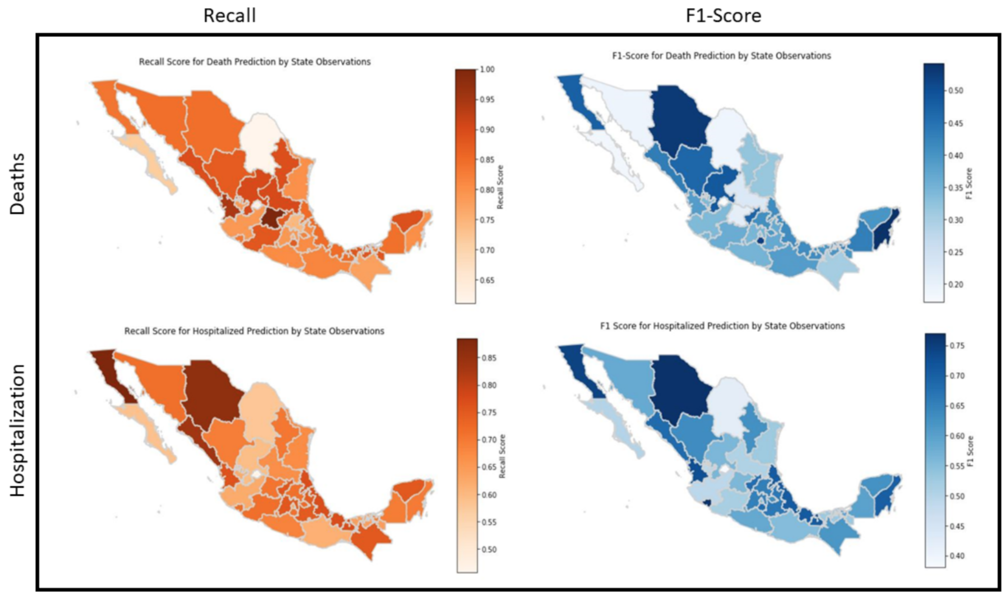


Figure 9: Relcall and F-1 scores for target variables by state


Recall
F1-Score
Deaths
Hospitalization
Recall Score for Death Prediction by State Observations
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Recall Score
F1-Score for Death Prediction by State Observations
0.50
0.45
0.40
0.35
0.30
0.25
0.20
F1 Score
Recall Score for Hospitalized Prediction by State Observations
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Recall Score
F1 Score for Hospitalized Prediction by State Observations
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
F1 Score