PROJECT PROPOSAL

1. *Project Title and Team*

Project Title: *COVID-19 in Mexico: Identifying vulnerable areas with an ML approach.*

Team Members:

- Roberto Barroso-Luque (barrosoluquer)
- Luz Stephanie Ramos Gomez (stephanieramos)
- Oscar Enrique Noriega Villarreal (onoriega)
- Jesica Maria Ramirez Toscano (jramireztoscano)

4-person Team Justification: We want to use this project to inform Mexican authorities and academia about the future impact of COVID-19 on Mexico. Hence, it is our responsibility to develop an adequate and relevant analysis that gives truthful information about the potential problems that Mexico will face. This is why we need a team of 4, to work carefully on every detail of the project. Additionally, every member of the team has a high motivation to help our country (Mexico) in the face of a health and economic crisis.

2. *Project Summary*

- What is the problem?

The current pandemic caused by the SARS-CoV-2 virus represents an unprecedented challenge to the global economy. Herculean efforts have been made to analyze, predict, and model the socioeconomic impacts of the pandemic to guide relevant policy. However, the majority of this analysis has revolved around just a handful of countries. In particular the USA, China, Italy, Spain, UK, Germany, South Korea, Taiwan, and Singapore[1]. Some characteristics of virus spread and its socioeconomic impacts are most probably shared among a majority of countries. Nonetheless, 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[2], institutional-weakness, and large informal economic sectors, Mexico faces a gigantic challenge to contain the virus. Unfortunately, like many other developing countries[3], there is a dearth of analysis with respect to Mexico. With this project, we hope to shed some light on the public health crises currently unfolding in Mexico and, more importantly, provide data-driven policy recommendations to combat the pandemic. Ultimately we seek to develop a covid19 vulnerability index based on health, economic, and disease transmission data at the municipality level.

[1] E.g. On a quarantine model of coronavirus infection and data analysis. Vitaly Volpert, Malay Banerjee, Sergei Petrovskii.

[2] More information on OECD DATA Health Spending.

[3] As coronavirus hits Latin America, expect serious and enduring effects, The Brookings Institution, Charles T. Call.

- Why is it an important policy problem? (Provide citations if possible to support your claims of relevance. Examples could be government requests for comments, white papers by policy organizations, etc.)

  The World Health Organization declared a global public health emergency on January 30th, yet despite the institution's warnings, political leaders failed to act in a timely fashion.
  Now, about 4 million people have been infected, and more than 200,000 have died from the COVID19 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 federal government anticipates a GDP decline of about 4%, while private institutions such as Credit Suisse expect a drop of 9.6% in GDP[4]. 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. The purpose of this project is to fill this analysis gap to identify vulnerable communities and speed up the governmental response.

- Who is the audience for your report?

  This report is meant to inform Mexican politicians and policy leaders. Our goal is to equip Mexican regulatory authorities, in particular public healthcare institutions such as the IMSS (Mexican Institute of Social Security), with analytical tools to predict most at-risk areas. 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 the country.

- What kinds of actions could be taken based on your results, and who is equipped to take those actions?

  With the proper information and analysis, policy-makers will be able to allocate resources to the most vulnerable communities and in consequence ameliorate the socio-economic costs caused by COVID-19 transmission. Regulatory authorities will be best equipped to make decisions based on the data and analysis available. We hope that with the appropriate information, political

[4] See Mexico's Finance Ministry Report and Credit Suisse Report.

leaders will employ the appropriate health, economic, and social policies to protect those most at risk.

- How will you validate whether your results might be relevant to your intended audience?

  We have been communicating with individuals currently working on informing the Mexican response to the COVID19 pandemic. In particular, some of this project has been inspired by the work done by TukanMX and investigators at *Mexico Como Vamos.* At the culmination of our project, we will make all our analysis and code available on GitHub, allowing fellow data-scientists and researchers to make use of our work for further investigations. If our project proves successful we will ask individuals in the institutions mentioned above to review our work and propose any improvements.

3. Data

- Describe the data you have and the data you'll need to collect

The Mexican government, together with the Secretary of health and the national council of science and technology (CONACYT), provides daily data on COVID-19 cases.

- ➢ DATOS ABIERTOS COVID-19 provided by the Secretary of Health gives information on a host of variables including age, gender, pre-existing conditions, geography, and other information on individual cases.
- ➢ COVID Mexico Daily Cases compiled by CONACYT provides higher spatial resolution with individual cases at the municipal level.

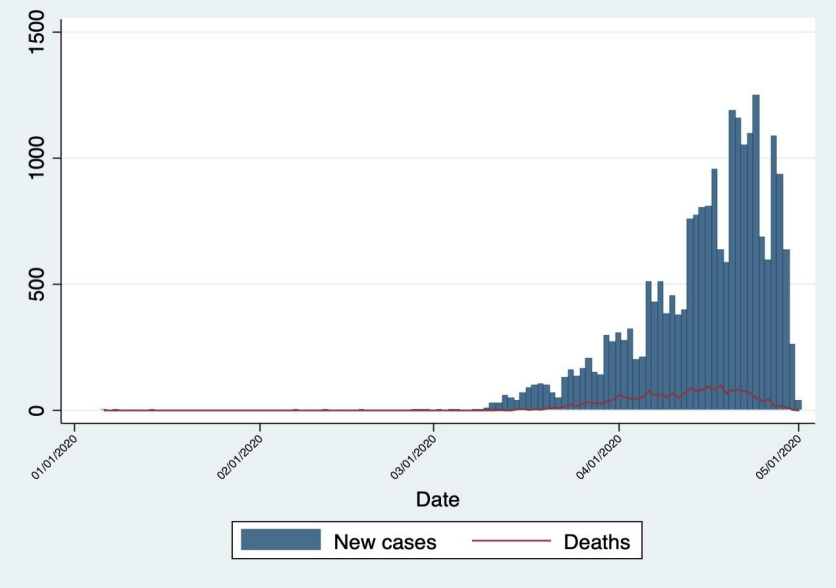
Furthermore, we plan to complement this data with the following:

- ➢ Economic Household Data made available by the National Institute of Statistics and Geography (INEGI) which provides information on both the informal and formal economic sectors.
- ➢ Health Resources Data provided by the Secretary of Health which contains information on the number of hospital beds and the number of doctors per geographic area. We hypothesize both of these variables will prove useful for our project.
- ➢ ENSANUT 2018 provides household-level data on health and nutrition indicators as well as associated social determinants.

- Include some descriptive stats that show you have enough to solve the problem.

Descriptive Statistics of COVID-19 in Mexico (up to May 2, 2020)	
Total Confirmed Cases	22,088
% female	41.87
% indigenous	1.4
% foreigners	0.96
% death	9.33
Suspected cases	14,536
Negative Cases	57,167
% cases with pre-existing conditions	51.58
Most Common pre-existing condition (%)	Diabetes (36.03%)
Mean Age	46.53
State with most cases (6,013)	Mexico City (27.22%)
State with least cases (29)	Colima (0.13%)
Average days between symptoms started and testing	4.21

Source: [COVID-19 Cases (Individual-level data) from the Secretary of Health of Mexico]



Source: [COVID-19 Cases (Individual-level data) from the Secretary of Health of Mexico]

- What analysis do you plan to perform or show on this data, to help inform your understanding of the data, as well as your design and selection of Machine Learning Models?

a. Clean daily COVID data sets and merge with relevant microeconomic and health indicators of each municipality in Mexico.
b. Visualize descriptive statistics among all potential predictor variables.
c. Visualize pre-existing conditions (diabetes, hypertension, etc) prevalence, related health metrics, and socioeconomic indicators through spatial analysis.
d. Inspect linear relationships and correlations between feature variables and outcomes, choose the best predictors.

4. *Machine Learning*

- What type of machine learning problem is this? Are you developing a classification technique? Regression? Prediction? Clearly articulate the learning that your resulting models will enable.

  Our tentative methodology is to start with regression methods to identify and screen predictors. Based on the regression results we will engineer appropriate outcome variables and bin geographical areas based on a calculated risk score. Once we have a valid outcome variable to describe different levels of risk/vulnerability we will use classification methods to train models, using "good" predictors, to forecast the relative risk of severe outbreaks faced by districts. Given the complicated nature and continually unfolding nature of the pandemic, we will probably revise our methods to best fit our needs.

- What types of models will you apply? Justify your choice of models. Your considerations could include the nature of your dataset (types and nature of features, size of the data), the requirements for model training or testing (e.g., real-time classification), or any other considerations you might have.

  As described above we will start our analysis with ridge regression, as this approach has shown promise in similar studies.[5] Initial regression analysis will allow us to identify the best predictors to build multivariate classifiers. Given our lack of knowledge of classification algorithms, we will start by using logistic regression. Logistic regression will allow us to see if our chosen features are, in fact, accurate in predictors of our calculated covid19 risk index. Furthermore, given its simplicity, logistic regression will allow us to gain intuition into whether our analysis is using the correct outcome variable and adapt further analysis accordingly. We then plan to use more powerful classification methods such as support-vector-machines, neural networks, and tree based-algorithms. While it is hard to justify the use of these more sophisticated classification methods we hope future course-material will equip us to use them appropriately.

5. *Evaluation*

- Describe your process for evaluating the models.

[5] Prediction of Infectious Disease Spread using twitter: A Case of Influenza

We will randomly divide the dataset into three subsets: training set, validation set, and testing set. We will build the predictive model using the training set. We will use the validation set to assess the performance of the model, to tune the parameters, and select the best model. To avoid overfitting of the model during meta-learning, we will use different subsets of the training data with a k-fold cross-validation approach. Finally, we will assess and compare the performance of each model in the test data with standard error metrics (e.g. RSME, MAE).

- How will you validate the correctness of the models? Be as specific as possible about your evaluation techniques (e.g., out-of-sample errors, imbalanced training sets, etc.).

  To validate the correctness of our models we will use various metrics for classification problems: classification accuracy (number of correct predictions made as a ratio of all predictions), confusion matrix (breakdown of correct and incorrect classification for each class), F-score (measures a test's accuracy by balancing the precision and the recall).

6. Ethics

- Briefly discuss the ethical implications of your work. Will your models suffer from any kind of bias? Will this potential bias impact the fairness of any proposed solutions?

  Our biggest concern is the accuracy of the data on COVID-19 infections. In general, we only observe severe cases in the data, in Mexico (as in many places in the world) only people with harsh symptoms are tested. Such imbalances mean the official figures are far off from the true numbers and will, in all certainty, bias our results. Furthermore, due to the lack of access to certain rural regions in the country, we are also concerned that our analysis will ignore many vulnerable communities that are not recorded in the data. In this sense, our results and proposed solutions won't take into account communities with little or no data. Lastly, we are concerned about the possibility that our analysis finds correlations between demographic variables such as gender, age, ethnicity, and higher COVID-19 rates. Our concern is that these correlates might provide ill-informed excuses to stigmatize certain groups in the population.

Additional References

- Building COVID Vulnerability Index. Caprio et al. 2020
- Identifying regions at risk with Google Trends: the impact of Covid-19 on US labor markets. Bank for International Settlements 2020.
- Vulnerability in the Informal Sector in times of COVID. Tukan Mexico.
- Fiscal and Human Resources of the Health Sector in Mexico. Tukan Mexico.