# Understanding Suicide: A First Look into the World and how Mexico behaves among countries.

1st Oliver Alejandro Velázquez Flores
*Computer Science*
*Tecnologico de Monterrey*
Monterrey, N.L., México
a01380501@itesm.mx

***Abstract*—In this work, suicide factors are going to be analyzed with the purpose of understanding this social and health problem that affects all societies around the world, despite clear economic and political differences among them. There are still many different variables that we do not understand about suicide nor the possible characteristics that can help us in closing this gap, economic and political conditions can provide strong evidence in these types of analyses, but there is still missing a connection between these features and the psychological aspect of this problem. Nonetheless, this analysis will be done by applying the CRISP-DM methodology with emphasize in time series clustering analysis and using predictive and classification techniques on 57 different countries worth of data. Furthermore, as we would see, the scope of this work is to build a strong bases for future works that can expand the knowledge through economical or psychological dimensions, as we will see in the discussion and future work sections.**

***Index Terms*—CRISP-DM, Data Mining, FBProphet, GDP Percentage Growth, Social Psychology, Suicides, Time Series Analysis.**

## I. Introduction

Suicide, as many other mental health issues, is a serious problem in the world, according to the World Health Organization (WHO) 800,000 people die due to suicide each year, that the most vulnerable people have age between 15 and 19 years, and that there are many more suicide attempts than suicides, which are a key indicator that there is something wrong with someone [1].

The WHO exhorts global governments to take actions in order to find plausible solutions for this public Health issue. Even though, the solutions are affordable and reliable, what is not is the recognition of a potential suicidal attempt. There are some key demographic, age, and gender indicators that can narrow down the group of potential people, there is still not enough information or problem understanding that can predict at an individual level. One of the main issues at hand is that, historically, suicide-related data was insufficient and incomplete mainly due to social and religious beliefs [2], and it was not worth the analysis. Right now, the WHO managed to put together a reliable source of information from different countries around the world, which at least, solves one of many other problems when working with suicide.

### A. Related Work

Related work in suicide related research is varied and displays how attributes correlate with each other. For instance **The relationship between suicide rates and age: an analysis of multinational data from the World Health Organization** by Ajit Shah [3] express the correlational relationship between age and suicide rates, but without implementing a Machine Learning or Data Mining approach.

The **Worldwide trends in suicide mortality, 1955–1989** work by Vecchia, *et al.* shows a similar approach than that of Ajit Shah, but with a larger and older time period [4].

The **National intelligence and suicide rate: an ecological study of 85 countries** offers interesting ideas in the different usage of different social statistics, such as the divorce rate in each country [5], which could add more valuable information for this project.

In **A Suicide Prevention Program in a Region With a Very High Suicide Rate** the researchers provide a prevention program in specific regions through depression management [6], which is important to have consideration of, but it is only a go-to spot, rather than an inspiration source, but it is still worth mentioning, because there are ideas that could work for future work.

This work will be divided in sections. In section II *METHODS*, there is going to be an explanation of the method that was followed to process missing data an the overall given data set. Then, in section III*METHODOLOGY* there is going to be a description of all the CRISP-DM methodology steps in detail. One thing to notice is that in step *Data Cleaning and Preprocessing* there are several features extracted. This new data set is only for reference only for future work. Furthermore, in section IV *Results* there is a complete overview of how different countries are cluster regardless of geographic position. Then in section V *Discussion and Future Work*, I discuss the importance of the understanding of suicide and the further steps that should be followed in future researches. Lastly, in the *Conclusion* section VI, all the important aspects of this work.

## II. Methods

Data is the principal raw material in Data Science in which all works are based on. There is a need of minimum

requirements for the data to be consider as such, one of which is complete data. Although given the nature of how data is collected and the data flow increase in the last decades, it is more and more complicated to gather complete observations; this is also called having *missing data*. Suicide data sets are not exempt from this behavior in general, and in particular the data set in this work has considerable missing data [7].

The raw data presented missing values in GDP and Population Percentage Growth for some countries. For the completion of these missing data, I search in the WorldBank data repository as well as in online economic cites such as Country Economy and with this I solved the first missing value problem.

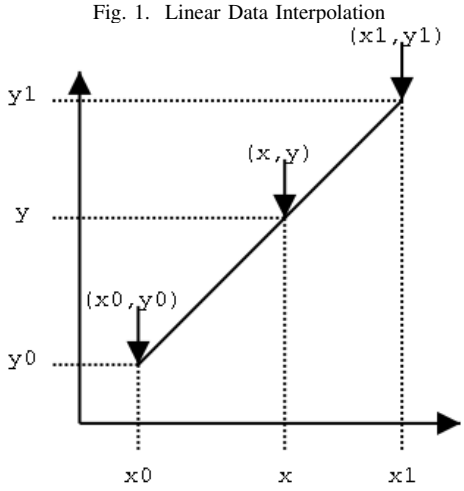
The second missing value problem is not shown until I did the Time Series Analysis. When setting the feature *Year* as index of the data, we can observe that there are several countries that did not have information for a given year or many, which case we dropped the countries that had not enough year-information to work with. For the rest, it is safer to apply data interpolation and reduce the bias that comes along with it.

### A. Linear Data Interpolation

For this second case, I applied equation 1 [8] to all the countries that had this problem. With 1 we can see what each $x$ represent.

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \tag{1}$$

Fig. 1. Linear Data Interpolation




One clear example can be seen in Fig. 2, It can be noticed that from Years 1985 to 1989 there are missing values. So the graph software just plots from 1990, but in Fig. 3, we can see that now, Germany has Data to work with. And, as we can see, the interpolation is not as accurate as it might be, but it is good enough to create clusters based on Time Series behavior.

One of the downsides of data interpolation is that it adds bias to any predictive or classifier model. One of the reasons that many countries were instead dropped.

Fig. 2. Data without Interpolation Example

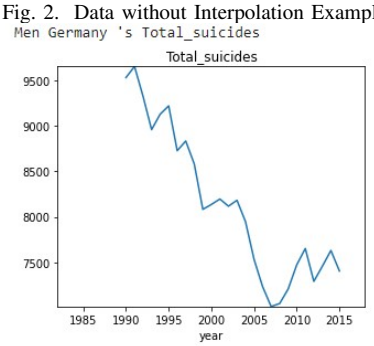
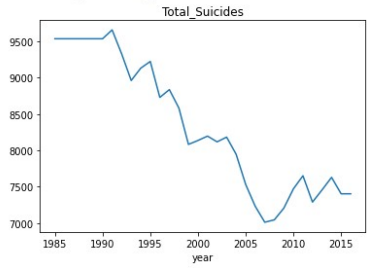



Fig. 3. Data with Interpolation Example




## III. Methodology

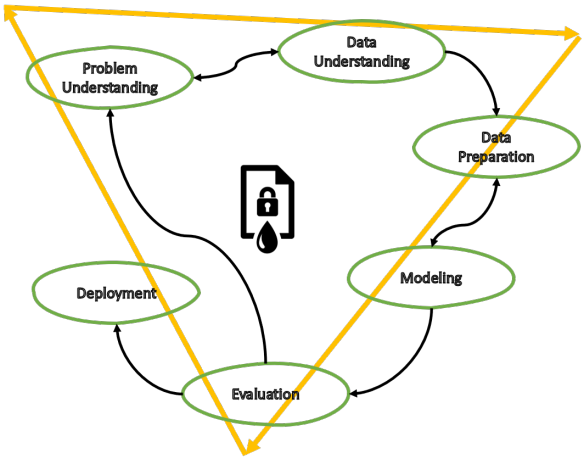
The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is frequently used when analysis a problem (usually business-related) in Machine Learning, Data Mining, and/or Big Data [9]. It should not be confused with the Software Development cycle; they could have similar structures, but their content is largely different. The key difference is the attention required in CRISP-DM, it iterates constantly and frequently among the process' steps, specially in *Evaluation* and the *Data Exploration* steps [9]. It was conceived formally in 2000 in Europe by Shearer, *et al.* [10], offering precise steps for problem-solving in Data-related organizations.

The steps included in the CRISP-DM methodology are: *Problem/Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, and *Deployment*, each step is essential in the Data Mining world, and provides clear criteria on how to solve specific problems. It is a iterative methodology, as mentioned before, and is widely used among researchers and businessmen alike.

### A. Problem Understanding

There is an implicit (and often explicit) necessity on understanding what are the factors creating the problem at hand and why is it important to be solved. These are intuitive reasons. That is why this step is often underestimated and skipped. Nonetheless, it should always be the first concern when solving a problem. One key factor is creativity, because it helps in figuring out a clever and easy-to-modify strategy that will make the subsequent steps easier to do [9]. There are two goals for this research, the first one is to generate a predictive

Fig. 4. A CRISP-DM Methodology diagram




model for the next five years from the last year in the data set, and the other one is to better understand the economic and social impacts on the increase (or decrease) of suicides.

It is also in the scope of this work to provide an answer to the following research questions:

a. Does historical events have an impact on suicide rates in countries around the world?
b. What is the age range in which suicide happens more often?
c. What are the countries that share the same trend in suicide?

### B. Data Understanding

This step is were the raw data is studied and analyzed. Data is usually found in a way that is not useful for the people that will use it, so an analysis of cost-benefit should be done before working with it, specially if there are missing values/data that could affect the overall process [9]. The data was gathered from the World Bank Group and World Health Organization repository and the Country Economy site for information on missing values in GDP annual growth on the dataset.

TABLE I
RAW DATA ATTRIBUTES

| Name | Type | Example |
|---|---|---|
| Country | Categorical | Mexico, Spain, ... |
| Year | Date (in years) | 1985, 1986, 1987, ... |
| Sex | Binary | Male, Female |
| Age | Categorical | 5-14 years, 15-24, ... |
| Suicides_no | Numeric | 0, 13, 2034, ... |
| Population perAgeGroup | Numeric | 363k, 1.997k, ... |
| Suicides/100k pop | Factorial | 55.65, 24.29, 12.37, ... |
| Country-year | Categorical | Argentina1985, ... |
| HDI for year | Numeric | 0.619, 0.656, 0.695 |
| GDP_for_year ($) | Numeric | 2,156,624,900, 2,126,000,000, 2,335,124,988, ... |
| GDP_per_capital ($) | Numeric | 769, 833, ... |
| Generation | Categorical | Silent, Boomer, Z, ... |

The raw data is laid out in Table I. These are the original attributes in the data downloaded from Kaggle [11], but the author thought that important data could be still added to the dataset (although he did not considered these attributes added to be part of a data extraction or selection process). The attributes added are shown in Table II. This is to have some tracking from year to year on the suicide tendencies among countries, specially with how GDP and population behaves across time.

TABLE II
NEW ATTRIBUTES ADDED TO THE DATASET

| Name | Type | Example |
|---|---|---|
| GDP_percentage_growth | Numerical | -5.189024352, 6.153377063, ... |
| Popuplation_percentage_growth | Numerical | 1.598562977, 1.584808921, ... |

The original data contains information of 101 countries for 32 years (from 1985 to 2016). Suicide numbers are divided in the six different age ranges available (5-14, 15-24, 25-34, 35-54, 55-74. 75+), as well in sex (female and male). There were countries that did not have enough information in said parameters and they had to be dropped from the analysis, because this missing data could not be obtained from the internet, nor it was a good idea to interpolate it; Now the country count is 57. One thing to point out, is that countries do not share the same time intervals. For instance, Mexico has available information from 1985 to 2015 (leaving 2016 empty) and Armenia's ranges from 1990 to 2003 and 2006 to 2016.

The attribute *HDI for year* (Human Development Index) could not be considered either due to missing data, so it was dropped from the data set.

### C. Data Cleaning and Preprocessing

Researchers and Engineers have different names for this step. It can also be called *Data Preparation* or *Data Processing*. Once the evaluation of the state of the data set is completed, the next step is to give shape and prepare the data in a way that will benefit the overall process; this means: select relevant data, fill missing values with statistical methods, create and extract attributes with the one in the raw data set, etc. At the end of this step, data should be more comprehensible [9].

For a Time Series Analysis, it is necessary that there is a one-to-one correspondence between the TimePeriod and the actual value to plot. Since this data set has many correspondences because of each age range and sex columns, the best way to avoid any inconsistencies, is to separate it first by gender and then by age range.

In this work, I decided to first analyze the problem only by gender, and then a combination of gender and age range; Meaning that, since we have 6 age ranges and two genders, then we will have twelve different data sets to work with, plus the two gender-only-data.

Now that evident irrelevant attributes have been removed, the next step is to fill missing data with statistical methods.

Although, there are serious concerns about the integrity of the model when interpolating data [12], it is a relevant method to apply when the benefits outweighs the cost of removing the missing values. If the author would have removed them, then he would only have five years to analyze, instead of 32. The methods applied were: *linear* and *pad* interpolations, the former with backward direction and the latter, forward, as explained in Section II-A.

New features were extracted from the original data set to complement and add completeness to the model. They are detailed in Table III. SPPI stands for Suicide-Population Index. It is the ratio between total number of suicide divided by total population per gender/sex (Eq. 2). This features are classified as statistical descriptive features.

$$SPPI = \frac{Total.No.Suicides(gender)}{Total.No.Population(gender)} \quad (2)$$

TABLE III
FEATURES EXTRACTED

| Name | Type | Extracted From: |
|---|---|---|
| Total_suicides | Numeric | Suicides_no, age |
| Total_population | Numeric | Population, age |
| Max_suicides | Numeric | Suicides_no, age |
| Max_population | Numeric | Population, age |
| Min_suicides | Numeric | Suicides_no, age |
| Min_population | Numeric | Population, age |
| Mean_suicides | Numeric | Suicides_no, age |
| Mean_population | Numeric | Population, age |
| SPPI | Factorial | Total_suicides, Total_population |

### *D. Exploratory Data Analysis*

For this section I will apply the TimeSeries Analysis, as well as clustering visualization. Here are some samples of TSA and Clustering visualization:

Descriptive statistical visualization is also a strong possibility.

This step, nonetheless, is for future work.

### *E. Modeling*

The *Modeling* phase is where computational and statistical models can be applied to our data, with the main objective of finding out different hidden and evident patterns to predict or classify within our solutions [9].

The twelve age range-gender data will be used only for TimeSeries and clustering analysis, whereas the gender-only data will also be used to generate Prophet models. The reason is that only this last set of data have enough observations to create a valid predictive model.

The first step is to see how every country behaves in similar fashion when considering economic and political factors, such as same government regime, system or so on.

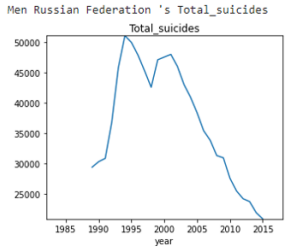
In figures 5, 6, we can see the behaviour of two former USSR countries: Russia and Ukraine. both have a heavy increase of suicides from 1985 to 1996. Another example can be seen in countries with capitalistic economies in figures 7 and 8.

After this modeling of singular countries, we can then proceed and apply clustering to all of the 57 countries in the data set. There were several criteria tested for clustering as well as the actual metric of them, as follows:

- Method, ward with Euclidean metric.
- Method, single with a Correlation metric.
- Method, single with a Spearman Correlation metric.
- Method, single with DTW metric.

For all the different age range and gender groups, the best method was Ward with Euclidean metric, and it is based on how balance they are (i.e., there are groups with more or less the same number of countries in them).

Fig. 5. Behavior of male Russia-Ukraine suicides



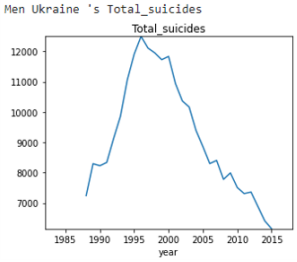




Fig. 6. Behavior of female Russia-Ukraine suicides


Fig. 7. Behavior of male USA-Mexico suicides

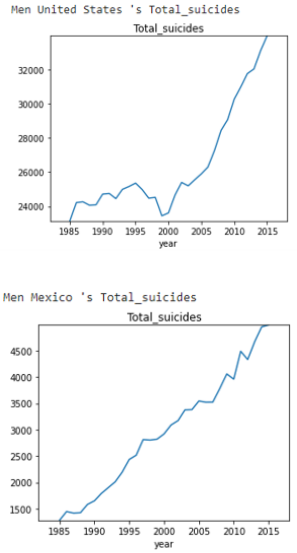
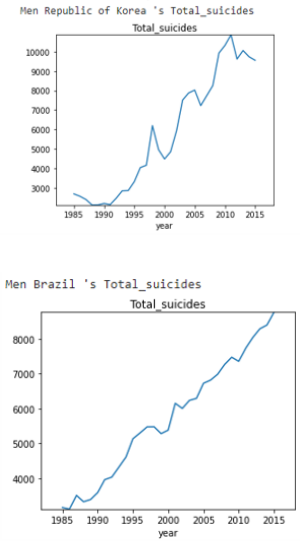



Fig. 8. Behavior of male Korea-Brazil suicides




For the group of Gender-Only separation I used The Facebook Forecasting tool Prophet. Prophet is an adapted forecasting tool that can be used as a .fit method, just like sklearn implements them, making it easier to use. This tool works well with outliers and trend in seasons in historical data, which works well with out type of data.

## IV. Results

In this section we can discuss the different clusters in which Mexico was placed. It is also worth noticing that the clusters that are going to be shown here are going to be mostly from men, due to the 4:1 ratio that men tend to suicide more than women.

In figure 9 We can observe that Mexico has similarities with Italy, Kazakhstan, Thailand, UK, Canada. This is considering all age ranges. We can then observe that Mexico is classified alone in one cluster in age ranges: 5-14 and 15-24, just like in figures 10 and 11. For the next age ranges we can see similarities with Mexico as follows:

- Age range 25-34: Kazakhstan, Korea, and Thailand, as seen in figure 12
- Age range 35-54: Australia, Italy, Romania, Spain, and Thailand as seen in figure 13
- Age range 55-74: Countries in figure 14
- Age range 75+: Countries in figure 17

We can see that there are different countries clustered based on age from different regions even.

For the case of the Prophet prediction, which can be done to any country in the data, but I'm going to show only for Mexico due to space.

We can see that the models are consistent with what the historical data shows of them.

### A. Evaluation

The *Evaluation* stage is for validating that each previous step was done correctly and towards the solution of the problem described in the *Problem Understanding* step. Here is where any adjustments (if any) should be made. Another important purpose is to be sure that our models are generalizable to other related problems [9].

One way of evaluation of this given methodology is to verify on each cluster and identify what do each country in it have in common. The interesting part of this analysis is that, first one needs to understand that there are many hidden variables that could indicate similarity between countries.

For the case of the predictive models, we can corroborate that all the data are within expectations when it becomes reliably public.

## V. Discussion and Future Work

This work presented an overview of how suicide is lurking in the data-available countries all over the world in the years from 1985 to 2016, taking into consideration GDP, annual GDP growth, age, gender, population, and annual population growth. The next step is to study each region with their unique set of social and economic problems, i.e., study Latin America, North America, Europe, Southeast Asia, and so forth, with emphasis in the Americas. Now with the pandemic caused by COVID-19, this analysis could be replicated and see what were the effects in different demographics around the world caused by burnout and desperation times like this one. It could be the case that "apparent" logic could cause a biased opinion and we could found out that suicide rate actually went down, but that would be jumping to conclusions too fast. It is best to conceptualize this ideas in hypothesis form and work around the follow-up research questions that could arouse.

Suicide is one of the many topics that strikes society (any in the world) that are unclear to humanity and has been present all along our civilization's history. There are not clear patterns or demographic characteristics, or even the successfulness of a person that could predict suicides (The clear examples being Robin Williams in 2014 and Anthony Bourdain in 2018,

Fig. 9. Mexico All ages-Male

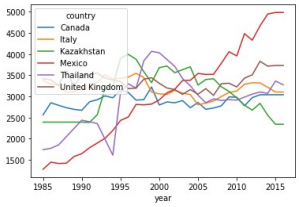



Fig. 10. Mexico 5-14age -Male

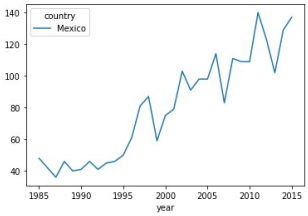



Fig. 11. Mexico 15-24age -Male

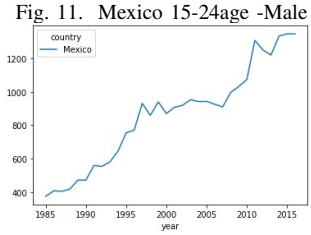
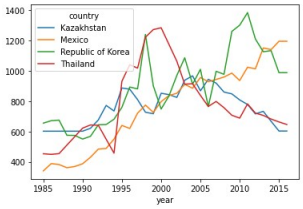



Fig. 12. Mexico 25-34age -Male




among many other actors and famous writers). There is still many work in this field in order to start grasping in actuality what does suicide mean and where it comes from.

As we could observed, historical events have an impact on suicide numbers going up. For instance, the Russian Federation and Ukraine (both part of the decayed USSR) showed a considerable increase in suicides in the next 10 years of the end of the Cold War in 1989 [13]. Another clear example is the Indian suicides in Latin America in the XIX and XX centuries in the sugar, coffee, and mines working camps, where they were forced to work on, with inhuman working conditions; they rather take their lives away than to work in said camps [14]. Although this was one of the research questions to answer in this work, it is just a glance on what the actual conclusions could be. It could be a correlated event, and it should be looked more in detail for future analysis.

## VI. Conclusions

### A. Deployment

This last step is where the models that we analyzed and built can be implemented in our business, organization, and/or project. In other words, that we can profit and can generate competitive advantage off of our solutions [9].

The predictive model is well suited for further years, and its performance will be better because it will account for newer tendencies in each country, as well as sudden changes in numbers.

From the result section we can observe that, there is not a single Age range that has higher suicide numbers than others. it depends on the country, region and year that is being analyzed. There should not be any conclusion that sets any age range as the most important in suicides, for that a complete analysis of the historical and economic problems in a given country has to be made.

One downside is that the predictive model has no way of identifying or predict any sudden change based on suicides numbers only, but any upward trend is.

Nonetheless, there are countries that are similar depending on age and gender, and adding new data to the predictive model should make it a more complete model with better understanding of the variables involved.

Fig. 13. Mexico 35-54age -Male

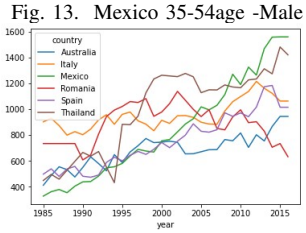



Fig. 14. Mexico 55-74age -Male

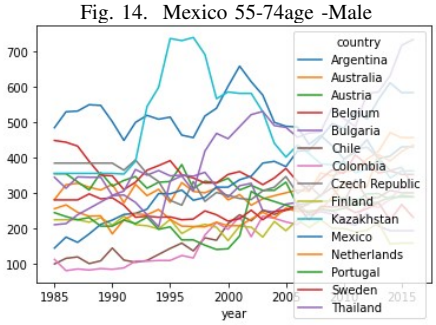
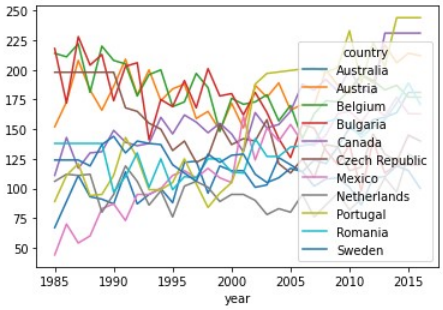



Fig. 15. Mexico 75age -Male





## ACKNOWLEDGMENT

I would like to thank my peer Ramón Hinojosa for his help and thoughts about my methodology and procedures, to thank Tecnológico de Monterrey and CONACyT (Consejo Nacional de Ciencia y Tecnología) for financial support


## REFERENCES


[1] W. H. Organization *et al.*, *Preventing suicide: A global imperative*. World Health Organization, 2014.

[2] J. M. Bertolote and A. Fleischmann, "Suicide and psychiatric diagnosis: a worldwide perspective," *World psychiatry*, vol. 1, no. 3, p. 181, 2002.

[3] A. Shah, "The relationship between suicide rates and age: an analysis of multinational data from the world health organization," *International Psychogeriatrics*, vol. 19, no. 6, p. 1141, 2007.

[4] C. La Vecchia, F. Lucchini, and F. Levi, "Worldwide trends in suicide mortality, 1955–1989," *Acta Psychiatrica Scandinavica*, vol. 90, no. 1, pp. 53–64, 1994.

[5] M. Voracek, "National intelligence and suicide rate: an ecological study of 85 countries," *Personality and Individual Differences*, vol. 37, no. 3, pp. 543–553, 2004.

[6] K. Szanto, S. Kalmar, H. Hendin, Z. Rihmer, and J. J. Mann, "A suicide prevention program in a region with a very high suicide rate," *Archives of general psychiatry*, vol. 64, no. 8, pp. 914–920, 2007.

[7] D. G. Altman and J. M. Bland, "Missing data," *Bmj*, vol. 334, no. 7590, pp. 424–424, 2007.

[8] S. C. Chapra and R. P. Canale, *Numerical methods for engineers*, vol. 2. Mcgraw-hill New York, 2011.

[9] F. Provost and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.", 2013.

[10] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[11] Rusty, "Suicide rates overview 1985 to 2016," 2018. [Online; accessed 27-April-2021].

[12] M. Belkin, A. Rakhlin, and A. B. Tsybakov, "Does data interpolation contradict statistical optimality?," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619, PMLR, 2019.

[13] S. J. Whitfield, *The culture of the Cold War*. JHU Press, 1996.

[14] E. Galeano, *Las venas abiertas de América Latina*. Siglo xxi, 2004.


Fig. 16. Mexico-Male Prediction

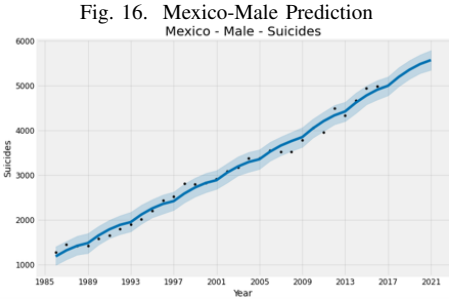



Fig. 17. Mexico-female Prediction