

Project Title

Team Members:

Monal Garg (PennKey: mgarg; Email: mgarg@sas.upenn.edu)

Amit Gupta (PennKey: ak Gupta; Email: ak Gupta@seas.upenn.edu)

Aashish Jain (PennKey: aashj99; Email: aashj99@seas.upenn.edu)

Moksh Jawa (PennKey: moksh; Email: moksh@seas.upenn.edu)

Assigned Project Mentor:

Jane Lee

Team Member Contributions:

Team Member	Contributions
Monal Garg	Past Work, Solution Approach, Machine Learning Algorithms, Visualizations, Report (continue if needed)
Amit Gupta	Problem Formulation, Algorithms Plan for Use, Machine Learning Algorithms, Visualizations, Report (continue if needed)
Aashish Jain	Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualizations, Report (continue if needed)
Moksh Jawa	Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualizations, Report (continue if needed)

Code Submission:

Public GitHub Repo can be found at <https://github.com/ak Gupta/CIS520-Final-Project>

Contents

1 Abstract 3

2 Introduction 3

2.1 Problem Motivation 3

2.2 Data Set and Source 3

2.3 Summary Statistics 3

3 Related Work 3

3.1 Related to Models 3

3.2 Related to Problem 3

4 Problem Formulation and Data 4

4.1 Data Processing Needed 4

4.2 Inputs, Outputs 4

4.3 Performance Measures 4

5 Solution Methods and Algorithms 4

5.1 Algorithms Used 4

6 Experimental Design and Results 4

6.1 Experiments 4

6.2 Results 4

7 Discussion and Reflections 5

7.1 Takeaways 5

8 Acknowledgements 5

9 References 6

10 Appendix 6

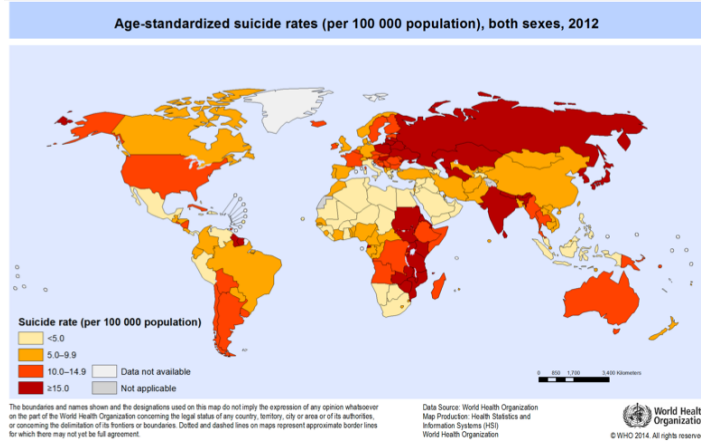
1 Abstract

Suicide is a major public health concern worldwide and varies from country to country. Because living in a different country and environment can greatly impact one's lifestyle, we were interested in collecting national data that could impact one's life over several years to better understand suicide rates internationally. With this data, we were interested in finding what factors could be possibly linked to certain suicide rate trends. To do this, we ran k -nearest neighbors, linear regression, and neural networks on the data... **TODO**

2 Introduction

2.1 Problem Motivation

Unfortunately, it is a sad reality that many individuals each year commit suicide. We in the U.S. are slowly being aware of this issue, but we wanted to understand suicide rates internationally. Especially, we wanted to look at these rates as they related to common features of each country, such as GDP, unemployment, and alcohol consumption. At a high-level we wanted to draw out any meaningful relationship between these macro features and suicide rates.



2.2 Data Set and Source

Our project grew from an intriguing World Health Organization data set with well over 35 years of suicide rates by country broken up by demographics.¹ In an effort to extend existing analyses though, we actually sourced global data from a variety of sources. This included a FiveThirtyEight data set on alcohol consumption². We also incorporated a World Bank Data Set on Youth Unemployment³ and another from the World Bank on population, fertility, and life expectancy⁴. A lot of the work done was with respect to time series and things like that. What we wanted to do, though, was take a look at combining data from multiple sources. Our project

team sought to understand a little bit more about the intricacies of sourcing data from multiple places. This proved to be a difficult challenge.

2.3 Summary Statistics

We've provided summary statistics for 4 of the datasets we will be using in our project shown in Tables 0, 1, 2, 3, some of which are located in the appendix.

Statistic	Average liters/person/year
Min	0
Max	14.4
Mean	4.71
Median	4.2
Standard Deviation	3.77

Table 0: Alcohol Consumption by Country

3 Related Work

3.1 Related to Models

Among the papers that analyzed economic factors, a regression model seemed to be the popular choice of method. For example, Pandey and Kaur paper used an Auto-Regressive Distributed Lag (ARDL) model⁵.

Most of the articles concerning the relationship of suicide rate and alcohol consumption related trends from a biological perspective utilized relatively simple statistical analysis⁶. Whereas medical research that involved more complex forms of data such as fMRI brain scans utilized algorithms such as Gaussian Naive Bayes⁷.

3.2 Related to Problem

During our research, we came across multiple research papers that analyze the correlation between suicide rates and either economical, behavioral, medical, or psychological factors. For example, the Pandey and Kaur paper investigates

suicidal trend and economic determinants in an Indian population⁸. Whereas articles published in medical journals explored suicide with alcohol/drug consumption, psychotic behavior, brain scans, etc⁹.

However, most papers tend to focus on either the economic or behavioral/medical factors. As suicide is likely governed by multiple factors, we seek to investigate this problem through a more holistic approach by simultaneously analyzing both medical, behavioral, and economic factors. Through this application of machine learning, we hope to gain insight on the relative role each factor plays in suicide and about the interaction of these factors.

Moreover, most papers perform a time series analysis. We provide a different approach by analyzing the data by country. Papers that do focus on certain regions usually provide analyses on one country or a specific region, whereas we will analyze global statistics.

4 Problem Formulation and Data

4.1 Data Processing Needed

Our data came in a variety of comma-separated formats. A lot of countries were labelled in differing manner and years were often a column and other times a point in a row. We used Excel Pivot Tables to make these more easily parsable and standardized. Later, blanks and other issues with inconsistencies are removed with Python to get an aggregated set of data. Other pre-processing includes removing blanks, reconciling some country names such as (Trinidad & Tobago vs. Trinidad and Tobago). **TODO**: talk about cleaning (Pandas, Excel)

4.2 Inputs, Outputs

Inputs are a variety of real-valued data points tied to the country. They are described above, but it remains to see which we might have to throw out due to excessive blanks. Most are also measured for a specific year, however some are just duplicated across all years (e.g. Alcohol Consumption). Outputs are a model, which given a countries characteristic features, can predict suicide rates. We have suicide rates by gender (as well as age demographics), so we may also seek a model which predicts a rate for male suicides and a rate for female suicides.

4.3 Performance Measures

TODO update to be what we actually used Given the regression nature of this problem, we believe a squared loss model is appropriate. **TODO**: address her comments

5 Solution Methods and Algorithms

TODO: update to what we actually did - why we choose, how they were implemented, what languages **TODO**: THIS WHOLE SECTION NEEDS TO BE REWRITTEN FOR WHAT WE ACTUALLY DID **TODO**

5.1 Algorithms Used

TODO describe algorithms, implementations, etc..
TODO We will start with least squares regression in unregularized, L_1 regularized, and L_2 regularized forms. We are not certain whether any of our features will be overly correlated so the unregularized least squares regression may fail. We also seek to explore Nearest Neighbor methods, but not with vanilla Euclidean distance, but with feature weighted scaling on the distances. Lastly, we will consider using decision stubs as features and feeding the transformed input space into another learning model, perhaps Neural Nets.

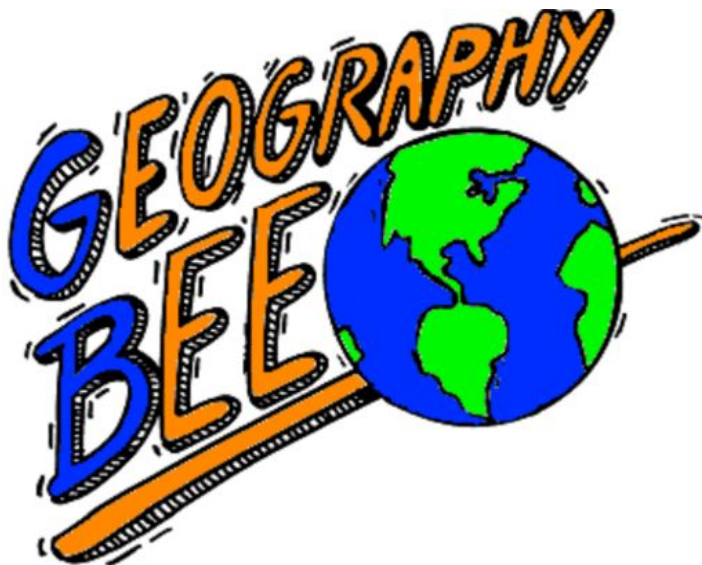
6 Experimental Design and Results

6.1 Experiments

TODO Describe how you set up your experiments, including choice of training/testing data, choice of parameters, etc, in a way that your results can be reproduced **TODO**

6.2 Results

TODO training/testing times, possible visualizations of results including both examples of good performance and examples of what types of errors are made, etc. **TODO**



The best predictor of suicide rates is the country label itself. Suicide rates in Ghana in 2013 are going to look like suicide rates in Ghana in 2012 most. Modelling this regression outcome as independent draws disregards the time series data and correlation of examples. As such, nearest neighbors performs exceptionally well, since it just recognized the country through these 8 input features.

7 Discussion and Reflections

TODO: Discussion about any sort of data specific things we gleaned, I already did the meta analysis **TODO**

7.1 Takeaways

This project was certainly an adventure in machine learning. Let us recall some of the lines from what we set out to do. Note, these were established a priori in our proposal: *“at a high-level we wanted to draw out any meaningful relationship between these macro features and suicide rates”* and *“we hope to gain insight on the relative role each factor plays in suicide and about the interaction of these factors”*. We were thinking that a model would illuminate which features were

coupled with suicide rates. As the project continued though, as noted in our Results section, models were essentially using the features to identify individual countries. We are careful in this report to not conflate correlation with causation and, as such, have left out implications of this country-recognizer model built.

Reflecting on our goal, it seems the problem we have chosen to tackle doesn't really make sense from a machine learning perspective. Sure, the factors we have chosen can be used to generate a model which fairly reliably outputs suicide rates. However, this correlation should not be taken to imply causation or even anything close to that. From a model story perspective, suicide rates are affected by many factors exogenous to our world of data.

When you consider some sort of underlying sampling distribution assumptions on the data, it doesn't make sense that suicide rates for countries on different continents with different socioeconomic statuses are coming from one distribution. Although this seems to say that this project wasn't an effective learning opportunity, we as a team feel quite the opposite. There are a finite number of countries in the world. In fact, there are 195 countries in the world. As such, the curse of dimensionality and other aspects of problems centered on larger data (think users of Netflix or game states) does not really factor in here. Had we as a group had more experience at the start of this project, perhaps we would have foreseen the uniqueness of countries leading to a model which could basically be a geography bee champion. However, we feel that the machine learning exercises carried out during the course of this project deepened our understanding of course algorithms, implications, and relevant considerations.

8 Acknowledgements

TODON/A

9 References

TODO: update?

Notes

¹<https://www.kaggle.com/szamil/who-suicide-statistics>

²<https://www.kaggle.com/fivethirtyeight/fivethirtyeight-alcohol-consumption-dataset>

³<https://www.kaggle.com/sovannt/world-bank-youth-unemployment>

⁴<https://www.kaggle.com/gemartin/world-bank-data-1960-to-2016>

⁵<https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/.08.pdf>

⁶<https://www.ncbi.nlm.nih.gov/pubmed/9229027>

⁷https://nocklab.fas.harvard.edu/files/nocklab/files/just_2017_machlearn_suicide_emotion_youth.pdf

⁸<https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/.08.pdf>

⁹<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872355/>

10 Appendix

	Youth Unemployed				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	0.6999	0.6999	0.5	0.6999	0.6999
Max	57.2	57.1	61.7	58	57.9
Mean	17.89	17.9	18.15	18.1	17.94
Median	14.9	14.52	14.4	14.1	14.12
Standard Deviation	10.54	10.88	11.43	11.67	11.55

Table 1: Youth Unemployment by Country

	Fertility Rates				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	1.06	1.11	1.16	1.13	1.21
Max	7.49	7.46	7.42	7.38	7.34
Mean	2.91	2.88	2.84	2.81	2.79
Median	2.41	2.37	2.34	2.34	2.33
Standard Deviation	1.45	1.42	1.39	1.37	1.34

Table 2: Fertility Rates by Country

	Life Expectancy				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	47.56	48.284	49.041	49.825	50.621
Max	82.9780488	83.4219512	85.4170732	83.8317073	83.9804878
Median	72.2783848	72.4719847	72.657	72.786	72.9707317
Standard Deviation	8.34929356	8.19102272	8.04819537	7.89104622	7.80484223

Table 3: Life Expectancy by Country