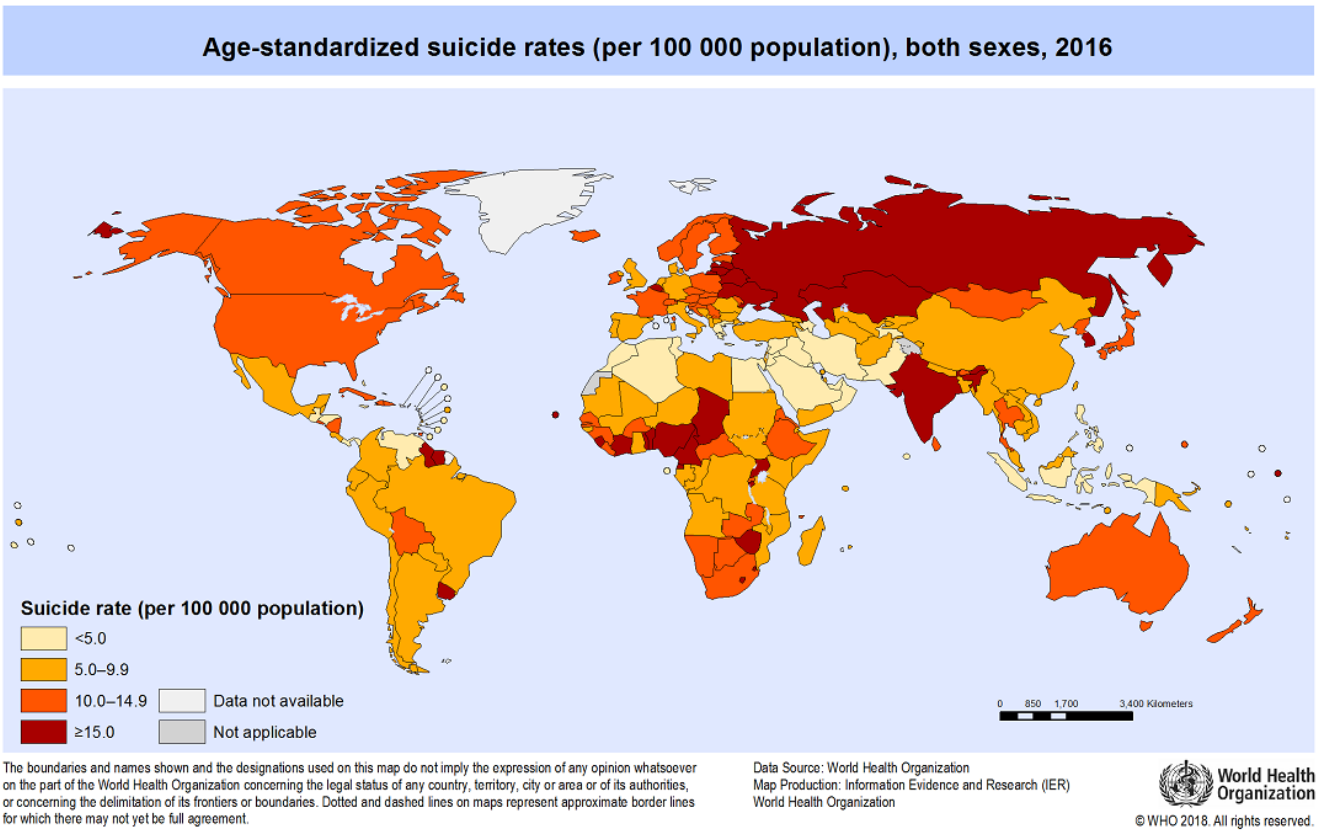
Machine Learning Engineer Nanodegree - Capstone Proposal

Li Chen Wu
January, 2019

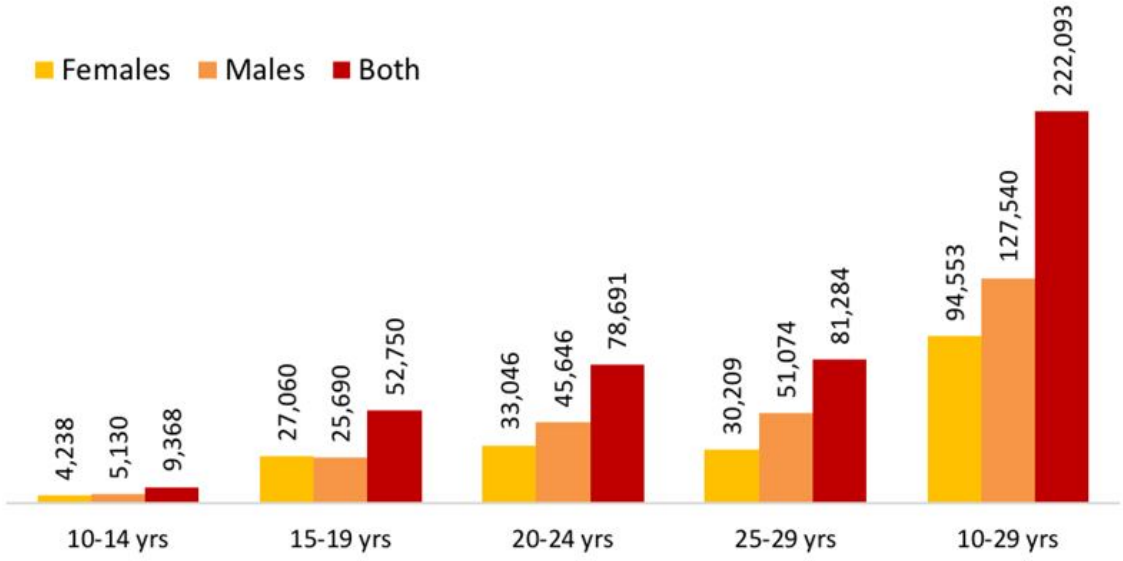
Domain Background

Nowadays, a lot of people from different countries die in suicide every year due to different reasons. According the statistics in Worth Health Orgnization (WHO) (https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ (https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/)), there are 8000 people die due to suicide each year, that is one person suicide in 40 seconds.

See the two figures below, it seems that the suicide rate is highly related with the areas, ages, and sex. The figures alsl show that high suicide rate usually occurs in some specific areas and ranges of age.

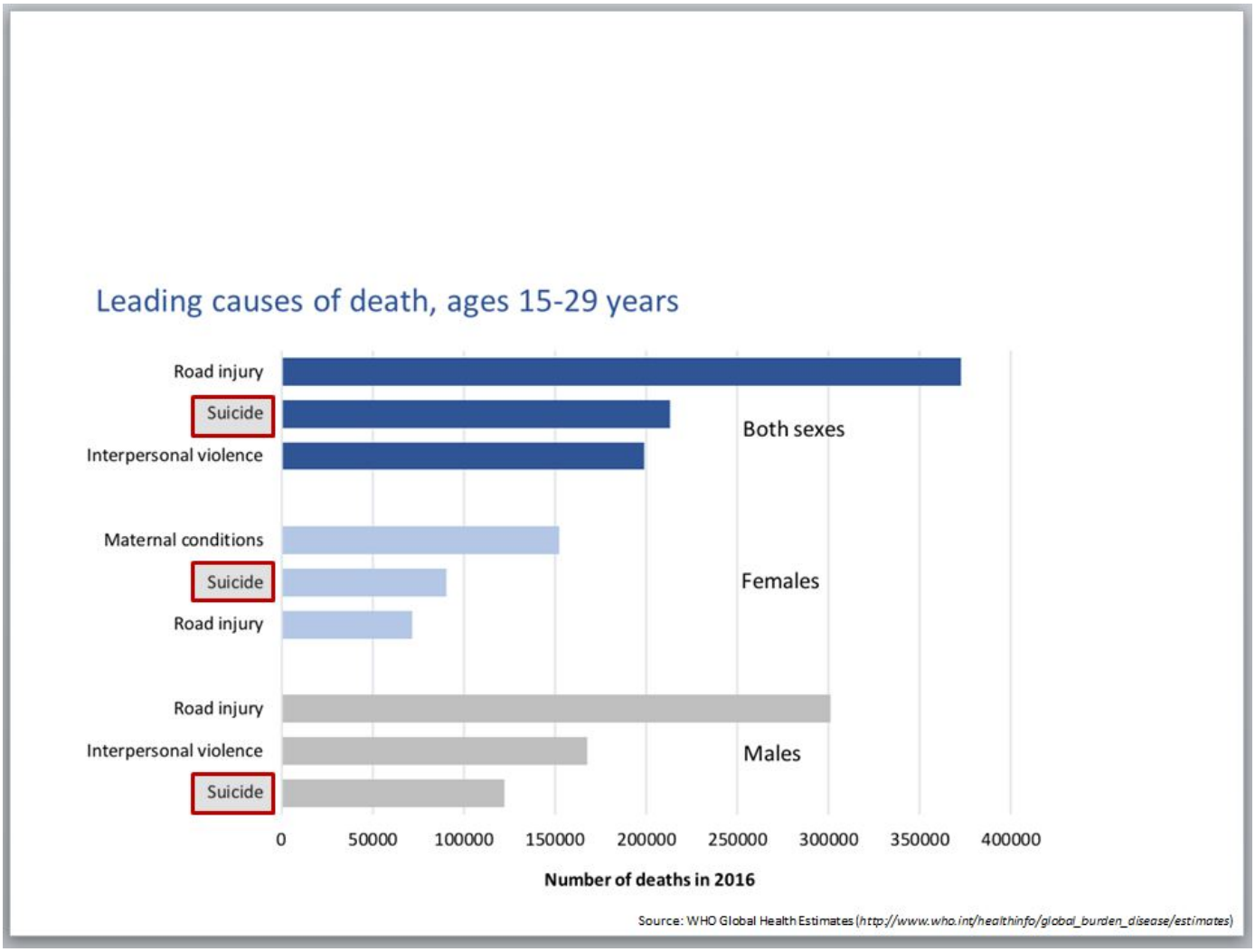


Number of suicides globally in young people, 2016



Source: WHO Global Health Estimates (*http://www.who.int/healthinfo/global_burden_disease/estimates*)

I was astonished of the huge number of suicide people and the suicide is also one of the leading cause of death in teenager.



There are many researchers do resarch on suicide analysis. Pete Burnap et al. [2015] use machine learning to classify the text on Twitter to find which are relating to suicide. Jihoon Oh et al. [2017] also use
artificial neural network classifier to detect the patients who has actual suicide attempts by the training data of patients' self-report psychiatric scales and questionnaires about their history of suicide attempts.

I think if we know a person who is tend to suicide beforehand, we can give them more help and avoid the suicide happening. Therefore, in this project, I would develop a classifier to predict a person who is prone to suicide or not.

Problem Statement

The goal of this project is to detect a person who is tend to suicide or not by input his/her basic information including country, sex, age, etc. After providing the informatoin to my model, my model will decide the person is tend to suicide by ouput Y or N.

Datasets and Inputs

The source of the datasets is from Kaggle WHO Suicide Statistics (https://www.kaggle.com/szamil/who-suicide-statistics). The data is the suicide numbers for each contries from 1979-2016 and the fields contains:

- country
- year: the year of the recode, 1979-2016
- sex: Male or Female
- age: Age group
- suicides_no: Number of suicides
- population: Number of all living people

Because there are no labels on the dataset, I would calculate the suicide rate and create the labels Y and N for each data. Hence two fields would be added to the dataset:

- suicides_rate: Suicide Rate of the country (suicides_no/population)
- label: Y (tend to suicide), N (not tend to suicide)

There are 43776 records in total in the dataset from 1979-2016 year. I think it is not necessary to refer the data which is too old, so I would use 2010-2016 as my dataset which is 7944 in total, and it remains 7044 after removing the missing data in population.

For spliting the data into training/validation/testing subsets, I would pick 80% for training/validation data and 20% for testing data in each year. Also I would pick the same ratio of labeled Y and N with the raw dataset.

The input fields will be a person's:

- country
- sex
- age

After input the fields above to the model, the model will output Y or N which indicate to whether tend to suicide or not.

Solution Statement

Data Preprocessing

Because the training data only provides the numbers of suicides and the population of the country each year, I would calculate the suicide rate for each country, and define the label is prone to suicide if the rate is higher than some threshold.

Algorithm

I would use supervised learning and train the model using
Decision Trees, Ensemble Methods, SVM, and so on. Comparison of these training algorithms and choose the best one.

Benchmark Model

For this problem, I define my benchmark model as logistic regression. I would try other solutions and compare to the result from logistic regression.

Evaluation Metrics

I would use AUC evaluation metric to evaluate the imbalanced binary classification model.

Project Design

Data Preprocessing

1. Import dataset which is download from kaggle.
2. Remove data of year from 1979 to 2009, only retain data from 2010 to 2016.
3. Remove the missing data of which population is empty and fill zero to suicides_no if the field is blank.
4. Calculate the suicides rate by suicides_no/population and add the result to the new filed.
5. Set the label Y/N for each data to the new filed.
   Threshold the suicides_rate, if the suicides_rate is higher than the threshold, mark the record as Y that is people tend to suicides, otherwise, mark it as N.
6. Because the field country, age group and sex are not numeric, I would do one hot encoding on them.
   The age in the dataset is grouped by 5 ranges, 5~14, 15~24, 25~34, 35~54, 55~74, 74+. Hence it is not necessary to do normalize on age, otherwise, I would apply one hot encoding on it.

Split Data

Split the dataset into three parts: 80% of the dataset for training, validation data and 20% for testing data in each year.

Model Training

I would try several supervised learning methods to training the model including decision tree, Ada boost and SVM, and then choose the best one as my result.

References

1. Worth Health Orgnization: https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ (https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/)
2. Kaggle: https://www.kaggle.com/szamil/who-suicide-statistics (https://www.kaggle.com/szamil/who-suicide-statistics)
3. Machine Classification and Analysis of Suicide-Related Communication on Twitter https://orca.cf.ac.uk/76188/1/p75-burnap.pdf (https://orca.cf.ac.uk/76188/1/p75-burnap.pdf)
4. Classification of Suicide Attempts through a Machine Learning Algorithm Based on Multiple Systemic Psychiatric Scales https://www.frontiersin.org/articles/10.3389/fpsyt.2017.00192/full (https://www.frontiersin.org/articles/10.3389/fpsyt.2017.00192/full)