Suicide rates overview 1985 to 2016

Pegah Khazaei

July 2020

Supervisor: Dr. Hossein Hajiabolhassan

Course: Machine Learning

Shahid Beheshti University

Department of Mathematics and Computer Science

This is a compiled dataset which is pulled from four other datasets linked by time and place, and is constructed to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

This dataset was first published in 2018.

Data, Variable Details and Models

This dataset contains the following variables:

1. Country
2. Year
3. Sex
4. Age
5. Suicides no
6. Population
7. Suicides/100k pop
8. Country-year
9. HDI for year
10. GDP for year
11. GDP per capita
12. Generation

GDP per capita is a measure of a country's economic output that accounts for its number of people. It divides the country's gross domestic product by its total population. That makes it the best measurement of a country's standard of living. It tells you how prosperous a country feels to each of its citizens

HDI combining three dimensions: A long and healthy life: Life expectancy at birth. Education index: Mean years of schooling and Expected years of schooling

Numpy is a package in Python used for Scientific Computing, matplotlib.pyplot is a plotting library used for 2D graphics ; Pandas is the most popular python library that is used for data analysis.
The dataset is in form of a csv file containing data in the above-mentioned columns,
Our model follows Supervised Learning, which consists in learning the link between two datasets: the observed data X and an external variable y that we are trying to predict, usually called “target” or “labels”. Most often, y is a 1D array of length n_samples.
All supervised estimators in scikit-learn implement a fit (X, y) method to fit the model and a predict(X) method that, given unlabeled observations X, returns the predicted labels y.

Research Questions:

- Estimate the percentage of men and women suicide globally?
- In which year suicide number is higher among men and women?
- What is the distribution of suicides among the different age groups?
- What are the suicide rates among the different generations?
- Looking at suicide Vs population.
- Looking at Suicides trends overs years.

In this project we will explore the following models:

- Linear regression
- Logistic regression
- Decision tree
- Random forest
- SVM (Support Vector Machine)

K-means Clustering

The task is to cluster the countries into two groups - the ones with high number of suicides and the ones with low number of suicides.