Suicide Rate Prediction

Shreya Gopal Sundari
Instructor: Prof. Travis Millburn

PROJECT PROPOSAL

Objective:

Suicide is a serious public health problem. The World Health Organization (WHO) estimates that every year close to 800 000 people take their own life, which is one person every 40 seconds and there are many more people who attempt suicide. Suicide occurs throughout the lifespan and was the second leading cause of death among 15-29-year-olds globally in 2016.

Suicide does not just occur in high-income countries but is a global phenomenon in all regions of the world. In fact, over 79% of global suicides occurred in low- and middle-income countries in 2016. On average, in US there are 129 suicides per day.

The objective of this project is to predict the suicide rates using Machine Learning algorithms and analyzing them to find correlated factors causing increase in suicide rates globally. The dataset used is available in Kaggle.

Dataset Details:

The dataset is borrowed from Kaggle. This is a compiled dataset pulled from four other datasets linked by time and place from year 1985 to 2016. The source of those datasets is WHO, World Bank, UNDP and a dataset published in Kaggle. The details of the dataset are:

- **Number of Instances:** 27820
- **Number of Attributes:** 12

The below table defines attributes in the dataset:

No.	Attribute Name	Description
1	country	Name of country
2	year	Year of the incident: 1985 to 2016
3	sex	Gender: male or female
4	age	Range of age in years
5	suicides_no	Number of incidents
6	population	Population based on the age group, sex, year and country
7	country-year	Combination of country and year
8	HDI for year	Human development index (HDI) for year

9	gdp_for_year ($)	GDP of the country for the year
10	gdp_per_capita ($)	GDP per capita of the country for the year
11	generation	Generation of the person
12	suicides/100k pop	Number of suicides for 100k population

The generation feature includes the following categories:

- Generation Z: Born 1996 – TBD
- Millennials: Born 1977 – 1995
- Generation X: Born 1965 – 1976
- Baby Boomers: Born 1946 – 1964
- Silent: Born 1927 - 1945
- G.I. Generation: Born 1901 – 1926

Statistical Tests:

Test 1: To check the difference in suicide rates between male and female

Using independent sample t-test to check the difference in suicide rates between male and female. The hypothesis statements for this test are:

H_0: There is no difference in the suicide rates among male and female (Null).

H_1: There is difference in the suicide rates among male and female (Alternate).

Test 2: To find out the dependence of suicide rate on the age.

Finding out whether there is a dependence of suicide rate on the age using the Chi-Square test. The hypothesis statements for this test are:

H_0: Suicide rate and age are independent (Null).

H_1: Suicide rate and age are dependent (Alternate).