Application of machine learning techniques for suicide analysis and rate prediction based on socio-economic and demographic indicators

Adu Boahene Quarshie
Faculty of Engineering, Environment and Computing
MSc. Data Science and Computational Intelligence
Coventry University, Coventry. United Kingdom
quarshiea@uni.coventry.ac.uk

***Abstract*— In this paper, the main objective is to analyse the suicidal rates globally, based on the socio-economic and demographic indicators. Machine learning techniques and algorithms like XG boost, Random forest regression model, Bayesian ridge regression model, Ridge regression model, Lasso and Elastic Net regression model implementing the use of python programming language which utilised learning libraries, prediction results and technique performance. Multiple data cleaning techniques were used to solve the issue of missing data in the data set for the analysis.**
These machine learning methods were trained, tested and evaluated to give which model had the best performance. The suicide rates of the countries in the train data was then predicted for the countries.

Keywords— machine learning; suicide rates; poverty; regression; data; python; random forest; bayesian ridge; elastic net; lasso; XGBoost; Ridge regression; stacking

I. INTRODUCTION

In every year, there are about 800,000 reported cases of suicide globally. This information has further ranked suicide to be part of the highest-ranking causes of death, coming at 15th worldwide. Again, suicide has also ranked as the 2nd and 5th leading cause of death in young adults of age range 15-29 and 30-49 respectively. More on these statistics is suicide overtaking mortality in maternity as the leading cause of death amongst girls that are aged 15-19 globally.

When it comes to suicide, there are many factors that can facilitate it in various countries, but we focus on a very specific topic which is a country's socio-economic growth, which I believe can affect- in various ways, the suicide rates in a country. The socio-economic growth of a country has many indicators including HDI, GDP per Capita, GDP on a yearly basis, Standard of living, etc. All these indicators can help us study or observe the rate of poverty in a country.

It is no argument that when it comes to low-and middle-income countries, poverty is at a higher rate compared to high-income countries. Poverty is a very convoluted concept and its measurement can raise certain arguments. The correlation between poverty and suicide rates of countries can also be very hard to draw conclusions on since it is a well-established fact that suicidal behaviours stem from an individual level, where a person's mental health and other personal factors are taken into consideration. When it comes to the macro level, socio-cultural, contextual and economic factors aid in the role in the aetiology of suicide, a positive correlation between completed suicide and unemployment, and between and suicide economic crises.

This paper aims at using the dataset of 101 countries to generate the general suicide rate of the countries, train, test and predict the suicide rates of countries when certain indicators available in the features of the dataset is given.

This paper is organised in the following way:

- Chapter II – The Dataset
- Chapter III – Data Pre-processing
- Chapter IV – Data Analysis and Visualisation
- Chapter V – Feature Engineering
- Chapter VI – Machine Learning Algorithms and Prediction Results
- Chapter VII – Discussion and Conclusion
- Chapter VIII – Appendix
- Chapter IX – References

II. THE DATA SET

The data set used for this analysis was obtained from Kaggle.com, it a dataset compiled from four other sets which is connected by time and place and was built to determine signals correlated to increased suicide rates among different countries globally.
It contains 27820 instances with each having 12 attributes with 1 attribute having missing values. Each instance is a country's suicide per 100,000 population, year, suicide number, sex, GDP per capita, GDP per year, total population, country year, HDI for year, and generation. The data was collected from 101 countries in a span of 32 years starting from 1985 to 2016. However, some countries don't have their data starting exactly from 1985 and ending in 2016. Out of the 12 attributes for the instances, 3 are categorical, 7 are numerical and 2 are objects. These features comprise of 2 floats, 5 integers and 5 objects.
It was observed that the feature 'HDI for year' had missing values. The number of missing values is 19456 which leaves the feature with 8364 values. The missing values is about 70% of the expected value.

TABLE 1. THE DATASET FEATURES

No.	Description	Type	Categorical Value Range
1	Country	Object	
2	Year	Numeric	
3	Sex	Categorical	Male, Female
4	Age	Categorical	5 – 14 15 – 24 25 – 34 35 – 54 55 – 74 74+
5	Suicide Number	Numeric	
6	Population	Numeric	
7	Suicide_Rate_100k	Numeric	
8	Country Year	Object	
9	HDI for year	Numeric	
10	GDP for year ($)	Numeric	
11	GDP per Capita ($)	Numeric	

12	Generation	Categorical	Silent Boomers Generation X G.I. Generation Millennials Generation Z

III. DATA PRE-PROCESSING

Dropped features:

Country Year is dropped because the country year became a redundant feature since the year of the country's data was already a feature in the dataset. Also, the country year had no significance when it came to the project objective.

The feature HDI for year was dropped because it had missing values (19456) that made up to 70% of the expected total data input (27820). The average of the data that was not lost was not a fitting value for replacing the missing values because it will render our dataset biased.

Data Sorting:

The data was sorted according to years in order to make the training set more balanced when it is split into train, test and validating sets. This is due to the fact that not all the countries had their data collection starting from 1985 and ending in 2016.

Feature renaming and numerical features correction:

The GDP per Capita and GDP per year features had to be renamed because the way it was previously saved made it impossible to code with, in Python.

The GDP for Year feature had commas (,) separating the figures so it was recognised as strings but had to be changed to numerical values by removing the commas and letting python recognise it as numerical inputs.

IV. DATA ANALYSIS AND VISUALISATION

Global suicide analyses and increase and decrease rates

Globally, the 101 countries under study in this paper mostly had an increasing rate in suicide individually. Japan, The Russian Federation and The United States saw their suicide counts ranging from 800,000 to 1,000,000+ in the span of 32 years which places them at the top 3 respectively as countries with the highest suicide counts from 1985 to 2016. Oman, Saint Lucia, Fiji, Bahamas and Dominica were amongst the countries that had a very low suicide count from 1985 - 2016

In the span of 32 years, there are countries that had their suicide rates increasing and decreasing. In the study of countries that had the highest decreasing rates of suicide, there are Kiribati, Estonia, Latvia, Lithuania, Sri Lanka, Hungary, Slovenia, Russian Federation, Finland and San Marino. The trends seen in the suicide rates of these countries saw it reducing its suicide rates in the span of the 32 years which made them the top 10 countries with the highest decreasing suicide rates. Contrary, there are the top 10 countries that saw their suicide rates significantly increasing over the 32 years. These countries are Malta, Belize, Chile, Uruguay, Cyprus Suriname, Guyana, Republic of Korea, Montenegro and Bosnia and Herzegovina.

Considering the Global Suicide Rate per 100,000 of the population of a country, the suicide rates throughout the years saw the highest rate of suicide in 1995 with a rate of over 15.0 per 100,000 population and 2014 with the lowest global suicide rate which was about 11.0 per 100,000 population.

Suicide rate according to age groups and sex:

The trend in the suicide rates of men and women in 32 years shows that there was a high rate of suicide amongst men in all age groups than women. However, between the ages of 5 – 14, there was the highest suicide rates for women across all the age groups.

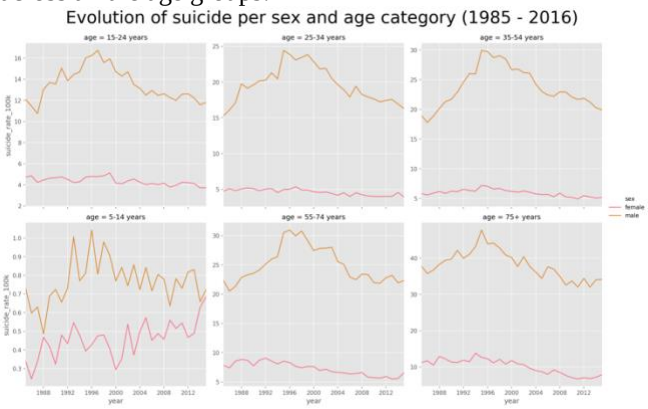

Fig 1. Suicide rates for sex and age category

Relationship between suicide rate and GDP per Capita:

It appears the direction of correlation changes with year with 1985 having an increasing trend. Although there doesn't seem to be a lot of data in 1985, as measured by the wider confidence bond.

While the next two decades don't show this trend or even slightly negative for 2005. 2015 appears to be slightly positive correlated.

Considering the representation, it is also observed that many low- and middle-income countries tend to have a higher or increasing suicide rates compared to higher-income countries or countries with higher GDP per Capita.

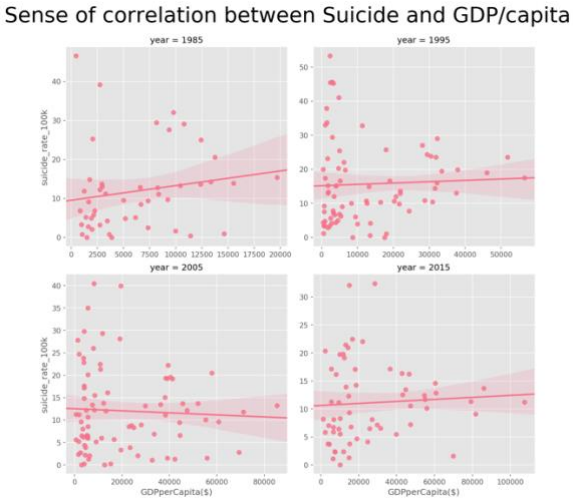

Fig 2. Correlation between Suicide and GDP per Capita in 3 decades

V. FEATURE ENGINEERING

Feature Creation:

In this paper, the suicide rates of countries are being explored. A new feature was created for the dataset to help with interpretation.

The feature would be simply named 'Suicide Rate' and it will be the percentage of suicide numbers over the population.

This will be the target variable (y) for our prediction after the models are fitted and trained.

$$Suicide\ Rate = \frac{Suicide\ number\ of\ country}{Population\ of\ country} \times 100$$

	country	year	sex	age	suicides_no	population	GDPforYear($)	GDPperCapita($)	generation	suicide_rate
10607	12	30	1	0	4.543295	13.417972	26.843650	10.665017	4	33.859773
4930	56	25	1	2	2.890372	10.940774	22.891297	10.004825	2	26.418348
7192	5	14	1	2	6.609349	14.820998	26.685289	9.999570	0	44.594495
4740	16	16	1	1	4.615121	13.252886	23.367710	7.525640	2	34.823513
12084	42	2	1	5	1.791759	10.976799	24.248483	9.260653	1	16.323151

Fig 3. Sample of dataset with the new suicide rate feature

Data Splitting:

The data set in order to be ready for model fitting is split into train and test dataset. The ratio for the splitting here is 80% to 20% respectively.

Upon exploring the 'suicide per 100,000 population' and the newly created feature 'suicide rate', we see that the average of suicide rate is 0.012% with the minimum rate being 0.0% and 0.22%. The suicides per 100,000 population is the same as the suicide rate created. Therefore, it is safe to drop the feature suicides per 100,000. This was converted to help with the interpretation in terms of percentage.

Feature Correlation:

The relationship between the new variable and the other variables were explored in the dataset and the correlation is represented below.

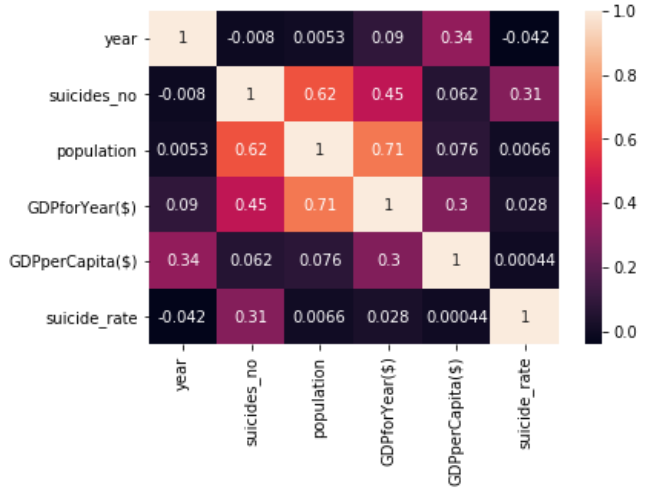

Fig 4. Heatmap of suicide rate variable's relationship with other variables

It can be seen that the other variables have a relationship with the variable of interest, that is suicide rates. The feature year is negatively correlated to suicide rate, thus as the year increases, there is a negative change in the suicide rates. This change due to an increase in year won't be so significant since the correlation is about -0.039.

Normalisation of highly skewed data:

Skewness is basically the measurement of the relative size of the two tails. Kurtosis is the measurement of the combination of sizes of the two tails. It takes measurement of the amount of probability that is in the tails.

The skewness and kurtosis of the target variable which is the suicide rate was explored for the train and test dataset, this was achieved by using SciPy-stats library utilising skew and kurtosis to analyse the dataset.

The skewness and kurtosis were found to be 2.94 and 11.7 respectively.

The dataset is highly skewed. This is because the skewness is greater than 1, that is, it is positively skewed (skewed to the right). for a normally distributed data, the skewness should be equal to zero.

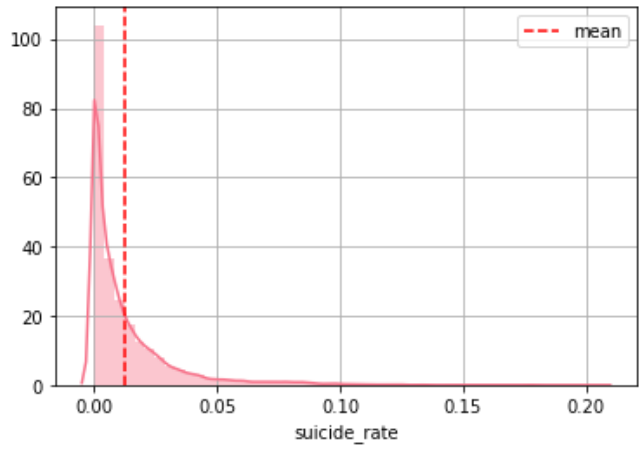

Fig 5. Skewed data of the suicide rate

In order to obtain a normal distribution, the dataset will be transformed by using the boxcox of it. This also deals with the outliers in the data. This was done using the python library SciPy-special where boxcox was imported to help with the transformation of the data.

The variable after this has skewness and kurtosis of 0.1 and - 0.39 respectively. The data values which were 0.0 were dropped.

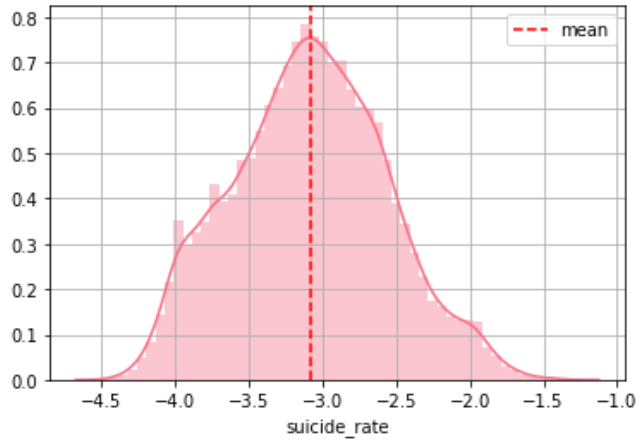

Fig 6. Transformed variable data

The length of the transformed values now sums up to 18804 and the total entry of the data was 22256. The new total data entry for the training set is 18804.

To perform the normalizing of highly skewed numerical variables and also to convert categorical or qualitative variables into numerical or quantitative variables of the entire dataset, we concatenate train dataset and test dataset.

The suicide rate feature was then dropped since it is the dependent variable or the target variable.

Label Encoding:

There are a large number of machine learning algorithms that only perform with numerical values but there are features in datasets that are categorical and can also be represented by text or strings.

Although categorical data can be represented by number, there are instances where machine learning algorithms would assign a high weight to a feature because it carries a higher assigned number.

To avoid such imbalance in the data, categorical features should be encoded to binary, quantitative or numerical values.

Here, we use the python library scikit-learn and importing LabelEncoder to change out categorical features namely;

country, year, sex, age and generation. LabelEncoder was used because in our data because the variables consist of two or more categories.

	country	year	sex	age	suicides_no	population	GDPforYear($)	GDPperCapita($)	generation	continent
0	89	29	0	0	3.610918	13.281003	27.075578	11.050207	4	EASTERN_EUROPE
1	64	29	0	2	3.951244	13.442988	26.936550	11.546786	2	C_W_OF_IND_STATES
2	70	11	0	2	3.496508	14.109013	25.532436	9.464052	0	ASIA
3	12	3	1	0	4.736198	13.522214	25.818045	9.770527	2	WESTERN_EUROPE
4	52	24	1	2	6.061457	12.995171	24.346023	9.426500	0	EASTERN_EUROPE

Fig 7. Sample of data after encoding

VI. MACHINE LEARNING ALGORITHMS AND PREDICTION RESULTS

The Regression models being implemented are the Random Forest Regression Model, Bayesian Ridge Regression Model, Ridge Regression Model, XGBoost Model, Lasso Regression Model and the ElasticNet Regression Model. The stacking method would later be implemented to see how it works to determine the better models to use.

Random Forest Regression Model:

A random forest is essentially a collection of decision trees. A random forest is an estimator which can fit a number of decision tree classifiers that are on various sub-samples of the dataset and utilises the averages to improve the predictive performance and can control over-fitting of the dataset. This is a good technique for our dataset since we will be dealing with a variable prediction [10]. (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

Bayesian Ridge Regression:

Bayesian ridge regression techniques can be utilised to include regularization parameters in an estimation procedure: the regularization parameter is not set in a hard sense but tuned to the data at hand [10]. (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

Ridge Regression:

Ridge regression can address some of the issues of Ordinary Least Squares by penalising the size of the coefficients [10]. (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

XG Boost:

Gradient Boosting can build an additive model in a forward stage-wise fashion; it essentially allows for optimization of arbitrary differentiable loss functions. In every stage n classes, regression trees are then fit on the negative gradient of the binomial or multinomial deviance loss function [10]. (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

Lasso Regression:

The Lasso is a linear model that can estimates sparse coefficients. It is very useful in some contexts because it tends to prefer solutions with few non-zero coefficients, which effectively reduce the number of features upon which the given solution depends[10]. (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

Elastic Net Regression Model:

Elastic Net is a linear regression model which is trained with ℓ1 and ℓ2-norm regularization of the coefficients. The combination of both allows for learning a sparse model where few of the weights are non-zero like Lasso, while it maintains the regularization properties of Ridge. (Scikit-learn: Machine Learning in Python, [10] Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

Comparing the models:

The best model is the model with the highest **R Squared** and the least error, that is, the root mean squared error. Therefore, from the results below shows that Random Forest Regression Model gives the highest **R Squared** and the least root mean square error.
Random Forest Regression Model is then used for the test dataset.

TABLE 2. MODEL FITTING RESULTS

Model	Score	R Squared
Random Forest	**0.015895228079430782**	**0.9998696615430251**
Bayesian Ridge	**0.08994744751939479**	**0.9706522974782487**
Ridged Regression	**0.08994801182637507**	**0.9706520475104724**
XG Boost	**0.039052233120954204**	**0.995751348430518**
Lasso	**0.0899803886489243**	**0.9706311615725988**
Elastic Net	**0.09000254962025801**	**0.9706169279125993**

Stacking:

To ensure that we get the best result for our feature prediction we introduce a method called stacking. This is to boost performance by combining weak models or improve on performance gain by strong models made up of different learning algorithms.
Stacking consists of building a meta-model that takes the output of base models as input.
Let's say you want to combine classifiers f1 and f2, both predicting the same set of classes. To
create a training example (ˆxi, ˆ yi) for the stacked model, set ˆxi = [f1(x), f2(x)] and ˆyi = yi.

What the algorithm does is to find the new train set and test set that best suit the models. It splits the data into four folds and works the algorithm to calculate the **R Squared** values of each fold. It combines all the six models and checks the Random Forest Regression each time. The algorithm then finds the full **R Squared** for all the folds and then splits the data according to all the models put together after the stacking. This gives us new data splits to use for the final train and testing to further help with the suicide rate prediction.

VII. DISCUSSION AND CONCLUSION

Although LMICs tend to have a lower suicidal rate, when compared to high-income countries (11.2 versus 12.7 per 100,000), 75.5% of suicides actually occur in LMICs. When you take the 10 countries with the highest suicidal rates, 8 out of the 10 are LMICs. Poverty, like suicide, is concentrated in LMICs, this is further proven in the graphical representation of the correlation of GDP per capita and suicide rates. The concentration of the of the countries were higher in the low-and middle-income countries.

The **R Squared** comparing the predicted variable and the true 'y' values was 0.7649593759965948. After the model stacking was introduced, the **R Squared** then changed to 0.7838641667769839. Comparing **R Squared** of just using the Random Forest Regression Model and the

stacking model, it can be seen that the stacking performs slightly better. This will make it the best to predict the suicide rate for our dataset.

Whenever GDP per capita, GDP per year, suicide number, age, sex and population are forecasted for a country, the suicide rates can be determined by the country with the help of this model which can help a country find ways to succeed in its suicide prevention programmes by finding ways to better itself in its economic growth.

```
prediction = pd.DataFrame()
prediction["suicide_rate"] = y_pred   #using the prediction of the stacking models
```

```
prediction.head(10)
```

	suicide_rate
0	0.020894
1	0.001006
2	0.002555
3	0.007132
4	0.016501
5	0.001239
6	0.000357
7	0.007179
8	0.033546
9	0.001466

Fig 9. Suicide Rate Prediction after Stacking

VIII. APPENDIX

Google Drive Link – Contains a **Jupyter notebook** of my code, the data set and a PDF copy of my coursework report: https://drive.google.com/open?id=1IXLQka4QVF2dEF7_zp577QRQitDO6QCc

IX. REFERENCES

1. Iemmi, Valentina, Bantjes, Jason, Coast, Ernestina, Channer, Kerrie, Leone, Tiziana, McDaid, David, Palfreyman, Alexis, Stephens, Bevan and Lund, Crick (2016) Suicide and poverty in low-income and middle-income countries: a systematic review. The Lancet Psychiatry, 3 (8). pp. 774-783. ISSN 22150366 http://eprints.lse.ac.uk/67387/
2. WHO. Preventing Suicide: A Global Imperative. Geneva: World Health Organization, 2014.
3. Petroni S, Patel V, Patton G. Why is suicide the leading killer of older adolescent girls? *Lancet* 2015; 386: 2031–2.
4. UN. Inequality Matters. Report of the World Social Situation 2013. New York: United Nations, 2013.
5. Alkire S, Foster JE, Seth S, et al. Multidimensional Poverty Measurement and Analysis. Oxford: Oxford University Press, 2015.
6. Stewart F, Saith R, Harriss-White B (eds). Defining Poverty in the Developing World. New York: Palgrave Macmillan, 2007.
7. Cooper S, Lund C, Kakuma R. The measurement of poverty in psychiatric epidemiology in LMICs: critical review and recommendations. *Soc Psychiatry Psychiatr Epidemiol* 2012; 47; 1499–1516.
8. Lund C, Breen A, Flisher AJ, et al. Poverty and common mental disorders in low- and middle-income countries: a systematic review. *Soc Sci Med* 2010; 71: 517–28.
9. Andriy Burkov, The Hundred-Page Machine Learning Book, 2019.
10. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
11. [R45f14345c000-1] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.
12. World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/
13. [Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook
14. World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#
15. United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506
16. 8 Das J, Do QT, Friedman J, et al. Mental health and poverty in developing countries: Revisiting the relationship. *Soc Sci Med* 2007; 65**:** 467–80.
17. 9 Patel V, Kleinman A. Poverty and common mental disorders in developing countries. *Bull World Health Organ* 2003; 81:609-15.
18. 10 McDaid D, Kennelly B. An economic perspective on suicide across the five continents. In: Wasserman D, Wasserman C (eds). Oxford Textbook of Suicidology and Suicide Prevention. Oxford: Oxford University Press, 2009.
19. 11 Durkheim E. Le Suicide. Etude de Sociologie. Paris: Félix Alcan, 1897.
20. 12 Henry AF, Short JF. Suicide and Homicide: Some Economic, Sociological and Psychological Aspects of Aggression. Glencoe: Free Press, 1954.
21. 13 Hamermesh DS, Soss NM. An economic theory of suicide. *J Polit Econ* 1974; 82: 83–98.
22. 14 Lester D, Yang B. Microsocionomics versus macrosocionomics as a model for examining suicide. *Psychol Rep* 1991; 69: 735–738.
23. 15 Marcotte DE. The economics of suicide, revisited. *Southern Econ J* 2003;69: 628–43.
24. 16 Fliege H, Lee JR, Grimm A, et al. Risk factors and correlates of deliberate self-harm behaviour: A

systematic review. *J Psychosom Res*, 2009; 66: 477–93.

25. 17 Hor K., Taylor M. Suicide and schizophrenia: a systematic review of rates and risk factors. *J Psychopharmacol* 2010; 24: 81–90.
26. 18 Schaffer A, Isometsä ET, Tondo L, et al. International Society for Bipolar Disorders Task Force on Suicide: meta-analyses and meta-regression of correlates of suicide attempts and suicide deaths in bipolar disorder. *Bipolar Disord* 2015; **17**: 1–16.
27. 19 Hawton K., Comabella CC, Haw C, et al. Risk factors for suicide in individuals with depression: a systematic review. *J Affect Disord* 2013; 147: 17–28.
28. 20 Colucci E, Lester D (eds.) Suicide and culture: understanding the context. Göttingen: Hogrefe Publishing, 2012.
29. 21 Platt S. Chapter 13: Inequalities and Suicidal Behaviour. In: O'Connor RC, Platt S, Gordon J (eds). International Handbook of Suicide Prevention: Research, Policy and Practice. Chichester: Wiley Blackwell, 2011.
30. 22 Vijayakumar L, John S, Pirkis J, et al. Vijayakumar, Lakshmi, Sujit John, Jane Pirkis, and Harvey Whiteford. Suicide in developing countries (2): risk factors. *Crisis* 2005; 26: 112–9.
31. 23 Nordt C, Warnke I, Seifritz E, et al. Modelling suicide and unemployment: a longitudinal analysis covering 63 countries, 2000–11. *Lancet Psychiatry* 2015; **2**: 239–45.
32. Are the Skewness and Kurtosis Useful Statistics? (2018) https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics