Classification of Suicide Data Using the Naïve Bayes Classifier Algorithm

Ankonisaho Nenungwi and Ritesh Ajoodha†

[1*]School of Computer Science and Applied Mathematics, Wits University,.

*Corresponding author(s). E-mail(s): ankonenungwi@gmail.com;
Contributing authors: ritesh.ajoodha@wits.ac.za;
†

Abstract

Suicide is the act of self-killing or ending one's own life[1]. One or more of Demographic, Social, Economic, and Psychological factors influence individuals to commit suicide. This study offers a new perspective into understanding attributes that make up the conceptual framework of suicidality through the Naïve Bayes algorithm. An in-depth exploratory data analysis followed by the Naïve Bayes classifier algorithm is used to categorize the number of suicides committed into Very-Low, Low, Medium, High and Very High. Using training data spanning from 1985 to 2016 in 101 countries and across 6 generations, suicide distribution is obtained. Classification is done to an accuracy level of 93.3465%.

Keywords: Suicide, Naïve Bayes Classifier, Classification, Exploratory analysis

1 Introduction

Death by suicide is labelled as premature because it often occurs before the biological end of an individual's life-span which is commonly said to be 75+ years. An important factor that is labelled responsible for suicides is psychological ache(psych-ache)[2]. About 800 000 individuals commit suicide each year[3]. Death by suicide, with the implementation of infallible preventative strategies can be combated and such high numbers avoided. The field of suicidology

would benefit from an improved understanding of factors that contribute to suicide attempts as well as solid prediction and/or classification models that shed more light into understanding patterns with which suicides occur. There are a number of underlying components making up the conceptual framework of suicidality. An understanding of these components is therefore crucial in advancements within the field of suicidology.

Age: Suicide is reported within the leading causes of death in Young adults (15-34)[4][5] [6]. This age-group consists of adolescents, college-students, individuals who are just starting out within the world of work, as well as those starting families of their own. The adolescent stage is associated with risky, impulsive behaviour, bullying amongst peers and some identity crises which can lead to suicidal behaviour[3][7]. The pressure that comes with being a college student, working hard to maintain an admirable grade point average as well as trying to maintain a social life can be too much for some individuals to bear leading to feelings of inadequacy and hopelessness which can push some individuals to commit suicide[8]. Each new phase of life comes with its own pressures and some individuals are unable to cope with these pressures and may decide that not living is the best out for them. The early-stage old age group (40-54) reported as the most at-risk often exhibit characteristics of a general sense of hopelessness and loneliness that come with the knowledge of impending death as one reaches old-age. Mass reporting of celebrity suicides is associated with an increase in suicides by reported method[9]. Such feelings can lead to suicide.

Gender: It is reported that men are twice as likely to successfully complete suicide as compared to women[10][6]. Men are in general more reluctant to seek-out help when they are facing mental health struggles thus making them more vulnerable to suicidal tendencies. Some methods of suicide are more like to be fatal than other methods. Men are reported to opt for more definitive methods of suicide such as hanging and the use of firearms as opposed to women who mostly use poisoning, a method which has a comparably higher survival rate[10].

Socioeconomic factors: In addition to demographic factors such as age, ethnicity etc, a country‘s economic standing, prevailing social and environmental changes as well as the region of origin contribute to the level of suicide risk[11][12]. Individuals from Low and Middle Income Countries are more likely to commit suicide as compared to those in richer countries[6]. Access to mental health care is in most countries a luxury that many cannot afford. Those in poorer countries may struggle to afford this mental health care and will most likely be alone in their struggles which may lead to worsening mental ill-health and individuals may resort to suicide. Increasing

economic growth and increasing unemployment rates are associated with an increase in female suicide rates, whereas decreasing economic growth, increasing divorce rates and increasing unemployment is associated with increasing suicide rates in males[11].

2 Methods

2.1 Exploratory data analysis

Exploratory Data Analysis(EDA) is a Data Science technique that analyzes data for the purpose of generating research questions as opposed to solving or answering pre-existing questions.[13]

Data Pre-processing: The Kaggle dataset recording suicide statistics between 1985 and 2016 is used for all analysis in this study. This dataset is a combination of four datasets linked through geographical and time indicators. The World Bank, World Health Organization and the United Nations Development Program are some of the listed sources of the data. The CSV file is converted to Excel. After this the dataset is loaded into Python Jupyter Notebook using the pandas library. Information about the dataset is extracted to reveal whether cleaning is necessary. A distinction is made that there are 12 attributes to be considered(in 12 columns). 11 of the columns have no null points. The HDI has only 8364 out of 27820 (30.1%) points. For the purpose of this study, the Human Development Index (HDI) will then be omitted. The *HDI for year* column is then dropped from the dataset. The information-seeking step is repeated and there are no null points, and all columns are accepted for the purpose of the study. The resultant dataset includes 101 countries under the column *country*, in the years *1985 to 2016*. The participants from each country are categorized in gender and age groups of *5-14,15-24, 35-54,55-74 and 75+* in years. Population size, GDP and GDP per capita, and generation are the demographical features recorded. Most important to the study is the number of suicides and number of suicides per one hundred thousand individuals of a specified sex and in a specified age-group.

Outlier Detection/Detect Abberation: Outliers are observation points that significantly differ from the majority of observations because of errors in the dataset or inherent variance within the data set.[14] Boxplots and scatter plots are the visual detection methods of choice for this study. In the boxplot, outliers are detected as points outside the whiskers of our box and whisker plot. For the scatter plot, we identify outliers as points that lie isolated from majority of points in the plot.

The **Z-score**: This measure is the number of standard deviations by which the value of a data point differs from the mean value of the data set. Outliers have z-scores greater than 3. The study calculates the zscore of each of the numerical columns. Once this is calculated, what follows is relating the zscores with the threshholds. This relation is then displayed in a line graph where it is clearly visible how many data points are within the normal range and which are outliers.

$$Z_{score}i = \frac{x_1 - \bar{x}}{s} \text{ where } X_i \ N(\mu, \sigma^2) \text{ and } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i - \bar{x^2}}$$

The Z-scores are based on that if X is normally distributed, then Z follows a standard normal distribution, and Z-scores greater than 3 are outliers[15].

The **Modified Z-scores** are much like the original Z-score method but shy away from using the sample mean and sample standard deviation s which are easily influenced by extreme value(s). This method opts for the *Median of the Absolute Deviation*.
$MAD = median|x_i - \tilde{x}|$, where $\tilde{x}$ is the sample mean.
$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$, where $E(MAD) = 0.675\sigma$ for large normally distributed data.

The **Median Absolute Deviation (MADe)**: This is a robust method that is **un**affected by extreme values.

Tukey's method: This method constructs a Boxplot using the IQR to filter out extreme data point values.states the applicability of this method to psychological data[16].
$LowOutliers = Q_1 - 1.5Q_3 - Q_1 = Q_1 - 1.5IQR$
$HighOutliers = Q_1 + 1.5Q_3 - Q_1 = Q_1 + 1.5IQR$

Outlier Treatment: (5)Winsorization is used to treat outliers.This method replaces data points that are above the 95^{th} percentile of the data with the 95^{th} percentile and any points below the 5^{th} percentile with the 5^{th} percentile. This method waters down the outlier. Outliers differ too much from the rest of the data, so in replacing them with conforming data points, they now look more like the rest of the data points[17].

Exploratory Analysis: Bar graphs are used to visually show the relationships between age, sex, and gdp against number of suicides.

2.2 Classification Using the Naïve Bayes Classifier

The **Naïve Bayes** Classifier is a simple and effective Classification algorithm which helps build fast machine learning models that can make quick predictions. It is a probabilistic classifier making predictions on the basis of the probability of an occurence. This method calculates the probability value based Bayes theorem and the algorithm is used to predict future probabilities based on experience from the training data set[18].

$P(w_k|v_j) = \frac{n_k+1}{n+|constant|}$ where :
$P(v_j)$ is the probability
$P(w_k|v_j)$ is the probability of w_k as class category v_j and
$|n_k|$ is the frequency of each category[19].

3 Results

The first part (EDA) of this study has examines possible features that affect suicide rates with reference to the conceptual framework of suicide consisting of age, gender and socio-economic factors.

Many features have an impact on suicide rates of countries. For example geographical features, people's religious beliefs, prevalence rates of substance use, sexual and physical abuse rates are effect the suicide rates. This could be the reason why there was no apparent link between suicide rates and the economic situation. Firstly, suicide rates are compared with the economic conditions of the country. These economic conditions are identified using the GDP and GDP per capita. According to the results of the analysis made, there is no apparent significant relationship between suicide rates and economic situation.

Studying the distribution of suicide rates according to age groups, it can be concluded that the methods used indicate that suicide rates are highest between the ages of 35 and 54.

A look into the distribution of suicide rates according to gender reveals that Suicide rates were in this case significantly higher in men than in women. This difference is maintained across all age-groups.

Data testing is done using the confusion matrix in order to get a more comprehensive result of the Naïve Bayes Classifier algorithm.

Confusion Matrix.

a	b	c	d	e	← Classified as
5764	89	2	0	208	a=low
685	9933	15	0	0	b=very low
0	0	5090	306	0	c=very high
0	0	161	2414	135	d=high
111	0	2	137	2768	e=medium

Using the confusion matrix to measure performance of the Naïve Bayes model, there are four possible outcomes representing the classification processes result.There can be a True Positive (TP), a True Negative (TN), a False Positive (FP), or a False Negative (FN). The True Positive (TP) reflects the percentage of elements correctly classified into a category they belong to. True Negative (TN) reflects the true amount of elements correctly classified outside a category. False Positive (FP) is the sum of all column values for each class, except for the TP value. False Negative (FN) is the sum of all row values for each class, except the TP value.

Detailed Accuracy By Class.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.951	0.037	0.879	0.951	0.913	0.889	0.988	0.929	low
	0.934	0.005	0.991	0.934	0.962	0.940	0.997	0.995	very low
	0.943	0.008	0.966	0.943	0.954	0.944	0.999	0.994	very high
	0.0878	0.019	0.836	0.878	0.856	0.841	0.993	0.937	high
	0.917	0.014	0.890	0.917	0.903	0.891	0.996	0.965	medium
Weigh Avg.	0.933	0.015	0.936	0.933	0.934	0.916	0,995	0,973	

4 Conclusion

In this study, gender, age, and socio-economic factors(country, gdp, and gdp per capita) can be used in suicide prediction. This research therefore proves the usefullness of the aforementioned attributes in suicidology. By using the Naïve Bayes Classifier algorithm, the very low class category is the one classified with greatest accuracy whereas the high class category is classified with the least precision. The Naïve Bayes Classifier algorithm has a very good accuracy of 93.3465%. [4]

References

[1] Wreen, M.: The definition of suicide. Health at a Glance. **14**(1), 1–23 (1988)

[2] Shneidman, E.S.: Commentary suicide as psychache. The Journal of Nervous and Mental Disease. **181**(3), 145–147 (1993)

[3] Jacob, J.: Conceptual framework - rings of suicide. International Journal of Recent Scientific Research. **9**(5) (2018)

[4] Sally C. Curtin, M.A.: State suicide rates among adolescents and young adults aged 10–24: United states, 2000–2018. National Vital Statistics Reports. **69**(11) (2020)

[5] Twenge, J.M.: Corrigendum: Increases in depressive symptoms, suicide-related outcomes, and suicide rates among u.s. adolescents after 2010 and links to increased new media screen time. Association for Psychological Science **6**, 3–17 (2018)

[6] Bachmann, S.: Epidemiology of suicide and the psychiatric perspective. International Journal of Environmental Research and Public Health. **15**(1425) (2018)

[7] Simon, G.E.: Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. American Journal of Psychiatry. **175**(10), 951–960 (2018)

[8] Holden, R.R.: Invalidating childhood environments and nonsuicidal self-injury in university students: Depression and mental pain as potential mediators. Wiley. **77**, 722–731 (2020)

[9] Niederkrotenthaler., T.: Association Between Suicide Reporting in the Media and Suicide: Systematic Review and Meta-analysis. BMJ, ??? (2020)

[10] Navaneelan, T.: Suicide rates: An overview. Health at a Glance. (82–624–X) (2012)

[11] Alfred Barth¨ Andreas Reiner, P., Robert Winker, M.: Socioeconomic factors and suicide: An analysis of 18 industrialized countries for the years 1983 through 2007. JOEM. **53**(3), 313–317 (2011)

[12] Xiang Liu, Y.L. Yi Huang: Prevalence, distribution, and associated factors of suicide attempts in young adolescents: School-based data from 40 low income and middle-income countrie. Plos One. (2018)

[13] Bogumil M. Konopka, M.O. Felicja Lwow: Exploratory data analysis

of a clinical study group: Development of a procedure for exploring multidimensional data. Health at a Glance. (2012)

[14] Jason W. Osborne, A.O.: Outliers: An evaluation of methodologies. Practical Assessment, Research, and Evaluation. **9**(6) (2004)

[15] K. Senthamarai Kannan, S.A. K. Manoj: Labeling methods for identifying outliers. International Journal of Statistics and Systems. **10**(2), 231–238 (2015)

[16] H. J. Keselman, J.C.R.: The tukey multiple comparison test: 1953-1976. Psychological Bulletin. **84**(5), 1050–1056 (1977)

[17] Dhiren Ghosh, A.V.: Outliers: An evaluation of methodologies. JSM., 3455–3460 (2012)

[18] Feng Xua, Z.Z.: A bayesian approach for predicting material accounting misstatements. Asia-Pacific Journal of Accounting Economics. **21**(4), 349–367 (2014)

[19] Hartanto, A.D.: Job seeker profile classification of twitter data using the naïve bayes classifier algorithm based on the disc method. 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). (2019)